Supplementary Information

Genome sequencing reveals insights into physiology and longevity of the naked mole rat

1 Genome sequencing and assembly

1.1 Genome sequencing

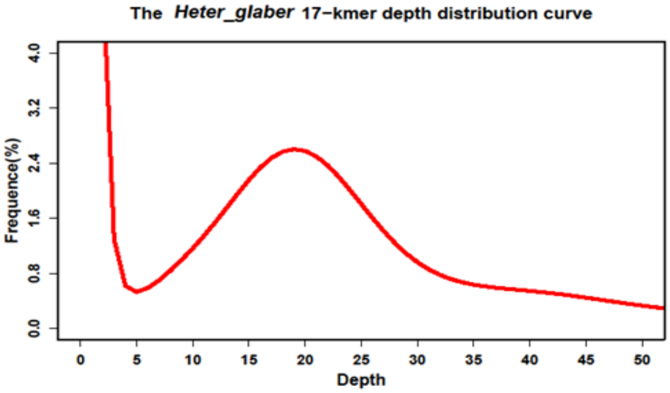
Animal experiments were approved by the University of Illinois at Chicago Institutional Animal Care and Use Committee. We used a whole genome shotgun strategy and the next-generation sequencing technologies on the Illumina HiSeq 2000 platform to sequence the genome of *Heterocephalus glaber* (the naked mole rat, NMR). DNA was extracted from an individual non-breeding male NMR. To decrease the risk of non-randomness, 18 paired-end sequencing libraries with insert sizes of 170 base pairs (bp), 350 bp, 500 bp, 800 bp, 2 kbp, 5 kbp, 10 kbp and 20 kbp were constructed and sequenced (48 lanes). In total, we generated about 475.78G of sequence, and following filtering out low quality and duplicated reads, 247G (90x coverage) was retained for assembly.

Supplementary Table 1. Parameters of genome sequencing of *Heterocephalus glaber*.

Pair-end libraries	Insert size	Total data (G)	Read length	Sequence coverage (X)	Physical coverage (X)
Illumina reads	170 bp	28.24	100	10.46	9.66
	350 bp	30.06	100	11.13	20.94
	500 bp	36.27	100	13.36	36.11
		5.38	150	1.99	3.98
	800 bp	23.46	100	8.68	37.78
		3.11	150	1.15	3.77
	2 kb	50.13	49	18.56	378.91
	5 kb	43.54	49	16.13	822.75
	10 kb	14.04	49	5.2	530.61
	20 kb	12.95	49	4.78	978.83
Total		247.18		91.55	

1.2 Estimation of genome size using k-mer

A k-mer refers to an artificial sequence division of K nucleotides iteratively from sequencing reads. A raw sequence read with L bp contains (L-K+1) k-mers, if the length of each k-mer is K bp. The frequency of each k-mer can be calculated from the genome sequence reads. k-mer frequencies along the sequence depth gradient follow a Poisson distribution in a given dataset, except for a higher representation of low frequencies due to sequencing errors, as sequencing errors affect the number of k-mers that may be orphan among all splitting k-mers. The genome size, G, was defined as G=K_num/K_depth, where the K_num is the total number of k-mers, and K_depth is the frequency occurring more frequently than other frequencies[1]. In the present study, K is 17, K_num is 52,143,337,243 and K_depth is 19; thus, the NMR genome size is estimated to be 2.74G, which is comparable to that of other rodents.



Supplementary Fig. 1. Seventeen-k-mer estimation of genome size. The genome size of NMR was estimated to be 2.74G based on reads from short insert size libraries.

Supplementary Table 2. *Heterocephalus glaber* 17-k-mer statistics.

Species	K	K_num	K_depth	Genome size	X
H. glaber	17	52,143,337,243	19	2,744,386,170	22.97

1.3 Genome assembly

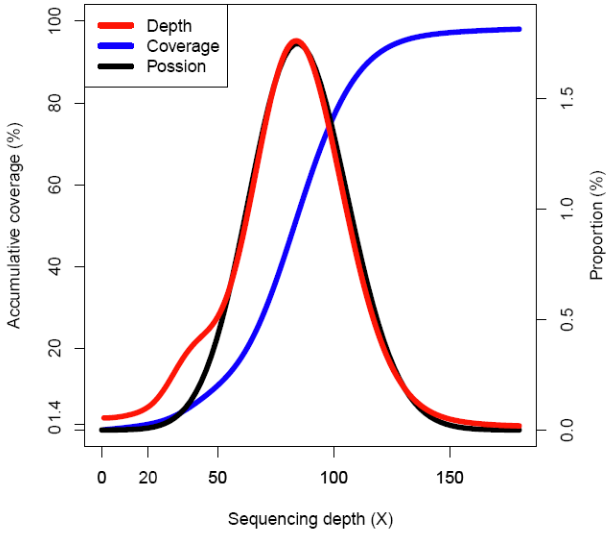
The NMR genome was assembled *de novo* using SOAPdenovo[1] (http://soap.genomics.org.cn). SOAPdenovo employs the *de Bruijn* graph algorithm in order to both simplify the assembly and reduce computational complexity. Low quality reads were filtered out and potential sequencing errors were removed or corrected by k-mer frequency methodology. We filtered out the following type of reads:

1. Reads having a ‘N’ over 10% of its length.
2. Reads from short insert-size libraries having more than 65% bases with the quality ≤ 7, and the reads from large insert-size libraries that contained more than 80% bases with the quality ≤ 7.
3. Reads with more than 10 bp from the adapter sequence (allowing no more than 2 bp mismatches).
4. Small insert size paired-end reads that overlapped ≥ 10 bp between the two ends.
5. Read 1 and read 2 of two paired-end reads that were completely identical (and thus considered to be the products of PCR duplication).
6. Reads having k-mer frequency <4 after correction (to minimize the influence of sequencing errors).

After these quality control and filtering steps, a total of 247G (or 91.5X) of the data were retained for assembly. SOAPdenovo first constructs the *de Bruijn* graph by splitting the reads from short insert size libraries (170-800 bp) into 41-mers and then merging the 41-mers; contigs are then collected which exhibit unambiguous connections in *de Bruijn* graphs. All reads were aligned onto the contigs for scaffold building using the paired-end information. This paired-end information was subsequently used to link contigs into scaffolds, step by step, from short insert sizes to long insert sizes.

About 126G (or 46.7X) of the data were used to build contigs, while all high quality read data were used to build scaffolds. Some intra-scaffold gaps were filled by local assembly using the reads in a read-pair where one end uniquely aligned to a contig whereas the other end was located within a gap. The final total contig size and N50 were 2.45G and 19.3K, respectively. The total scaffold size and N50 were 2.66G and 1.59M, respectively (Supplementary Table 3). To access assembly quality, high quality reads that satisfied our filtering criteria were aligned onto the assembly using BWA[2] with default parameters. A total of 97.4% reads could be mapped and they covered 99.7% of the assembly, excluding gaps. This observation suggests that nearly the entire NMR genome was represented in our assembly. However, with this information, we did not assess the case of collapsed regions (i.e., multiple copies of similar sequence in the genome, wherein not all copies were represented in the assembly). To test for completeness of the assembly, the sequencing depth of each base was calculated from the alignment, the proportion of a given depth was calculated, plotted, and compared to the theoretical Poisson distribution with a mean corresponding to the peak (here, it is 88).

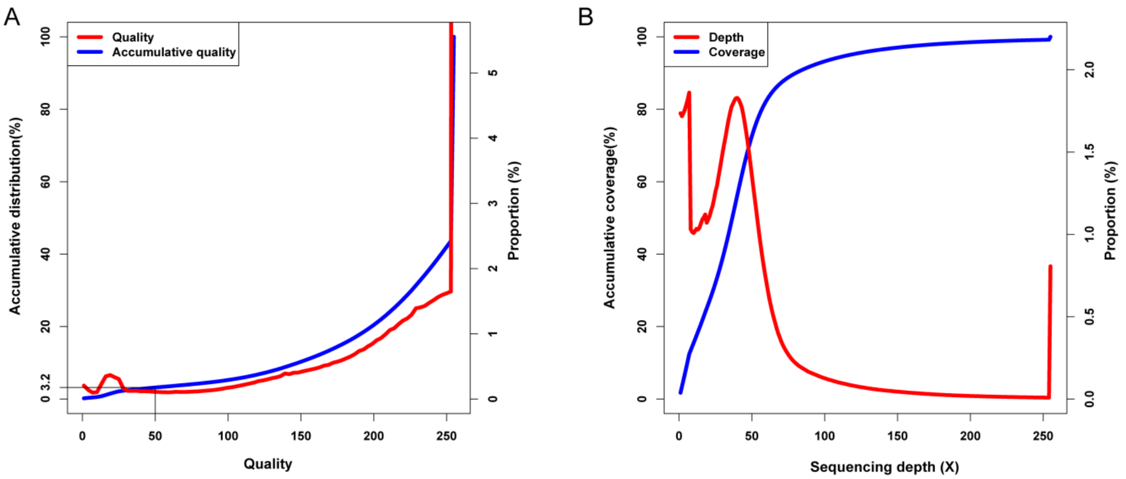
Supplementary Fig. 2 shows that the proportion of depth distribution fits a Poisson distribution, except for a shoulder, which appears from the heterozygous sex chromosome. If genomic regions collapse due to assembly quality, these regions would possess a higher than expected sequencing depth, i.e., if two copies are merged into a single copy, the depth of the assembled region is expected to be two-fold higher than the expected value. We found that approximately 1.9% and 1.0% of the assembly possessed a depth two or three fold higher than 90X, which is likely due to the highly similar repetitive sequences, such as centromere and telomere. Overall, we did not observe an unexpectedly high proportion of genomic regions with higher sequencing depth, suggesting that collapsibility is not a problem of our assembly. In turn, this observation suggests the occurrence of fewer repetitive sequences in the NMR genome. These results provide evidence for the completeness of our assembly. We also found that over 98.6% of genomic regions in our sequenced NMR genome were covered by at least 20 reads, and the inconsistent loci in the assembly were then corrected based on the uniquely mapped reads. Thus, high accuracy of the assembly at a single nucleotide level was obtained. The high density and gradient distribution of the distance of the paired-end information provide high confidence of the scaffolding. Discrepancy between the contigs/scaffolds and paired-end information may suggest mis-assembly or structure variation between two haplotypes. To further test for possible contigs being mis-joined into scaffolds, we analyzed paired-end information and found that more than 99.6% and 76.3% of paired-ends (where both ends could be uniquely mapped onto the assembly) were in the correct orientation and at the expected distance according to the utilized short and long insert size libraries, respectively. The proportion from the long insert library was significantly lower than that from the short insert libraries due to a cyclization step during long insert size library construction, which introduces DNA sequences with the size of approximately 500 bp instead of the expected length. When such paired-ends were excluded, the proportion increased to 97.0%. Overall, these tests suggested that the contigs and scaffolds were well consistent with the extremely high density of paired-end reads, which in turn indicated high quality of the assembly.



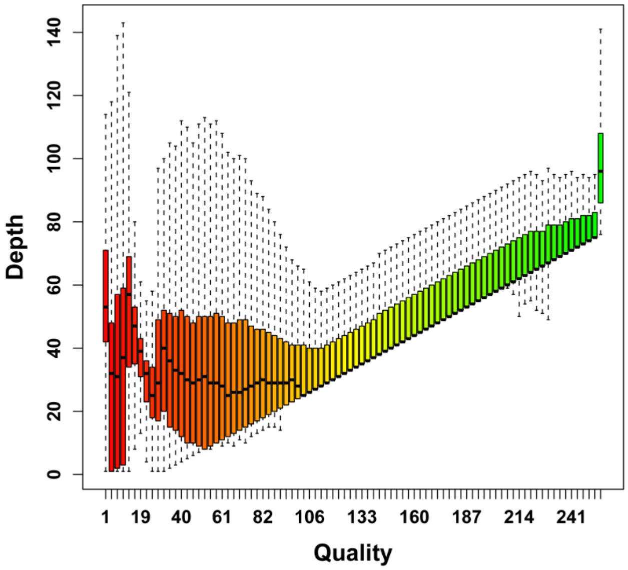
Supplementary Fig. 2. Sequencing depth distribution of the *H. glaber* genome. All high quality reads were aligned onto the assembly and the sequencing depth at each position was calculated. The red curve with a peak at 88 denotes the proportion of the genome in a given sequencing depth, while the blue curve show the accumulative coverage of the genome. A theoretical Poisson distribution with λ=88 is also plotted for comparison. The sequencing depth distribution fits well with Poisson distribution. Approximately 98.6% of the genome was covered by at least 20 reads.

Supplementary Table 3. Statistics of the assembly of the NMR genome.

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	4,762	131,974	330,812	1,861
N80	8,547	94,525	631,399	1,296
N70	11,981	70,462	923,861	948
N60	15,473	52,493	1,230,640	699
N50	19,307	38,321	1,585,568	508
Longest	178,884		7,787,482	
Total Size	2,448,567,728		2,664,766,285	
Total Number(>100 bp)		447,279		181,133
Total Number(>2 kb)		174,202		5,893



Supplementary Fig. 3. Distribution of consensus quality and sequencing depth with low quality. (A) The red curve denotes density distribution of a given consensus quality ranging from 0 to 255. The blue curve shows accumulative density distribution of consensus quality. (B) Depth distribution of the assembly with the consensus quality of < 50 displayed a peak at half of the whole genome (88-fold coverage), suggesting a relatively low assembly quality was with the lower depth compared to the whole genome.



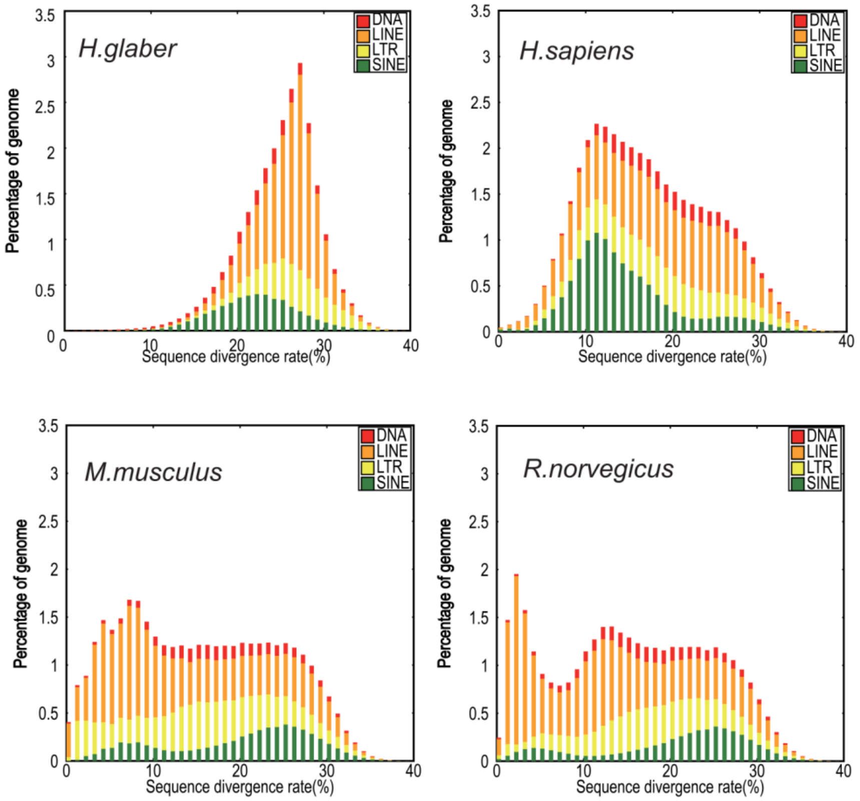
Supplementary Fig. 4. Relationship between consensus quality and depth. The X-axis denotes the quality inferred from a Bayesian model, and the Y-axis shows the sequencing depth inferred from short read alignments. This distribution shows a good correlation between consensus quality and depth.

2 Repeat annotation

Tandem repeats were searched across the genome using the software Tandem Repeats Finder (TRF)[3]. Transposable elements (TEs) were predicted in the genome by homology to RepBase sequences using RepeatProteinMask and RepeatMasker[4] with default parameters. For better comparison with other mammals, we employed the same pipeline and parameters to re-run the repeat annotation in human, mouse and rat genomes as shown in Supplementary Table 4. The diversity distribution of the detected TEs, compared with consensus sequences derived from Repbase, revealed that NMR had a relatively high diversity compared to the other three genomes in all four classified TEs (Supplementary Fig. 5).

Supplementary Table 4. TE comparison in NMR and other mammalian genomes.

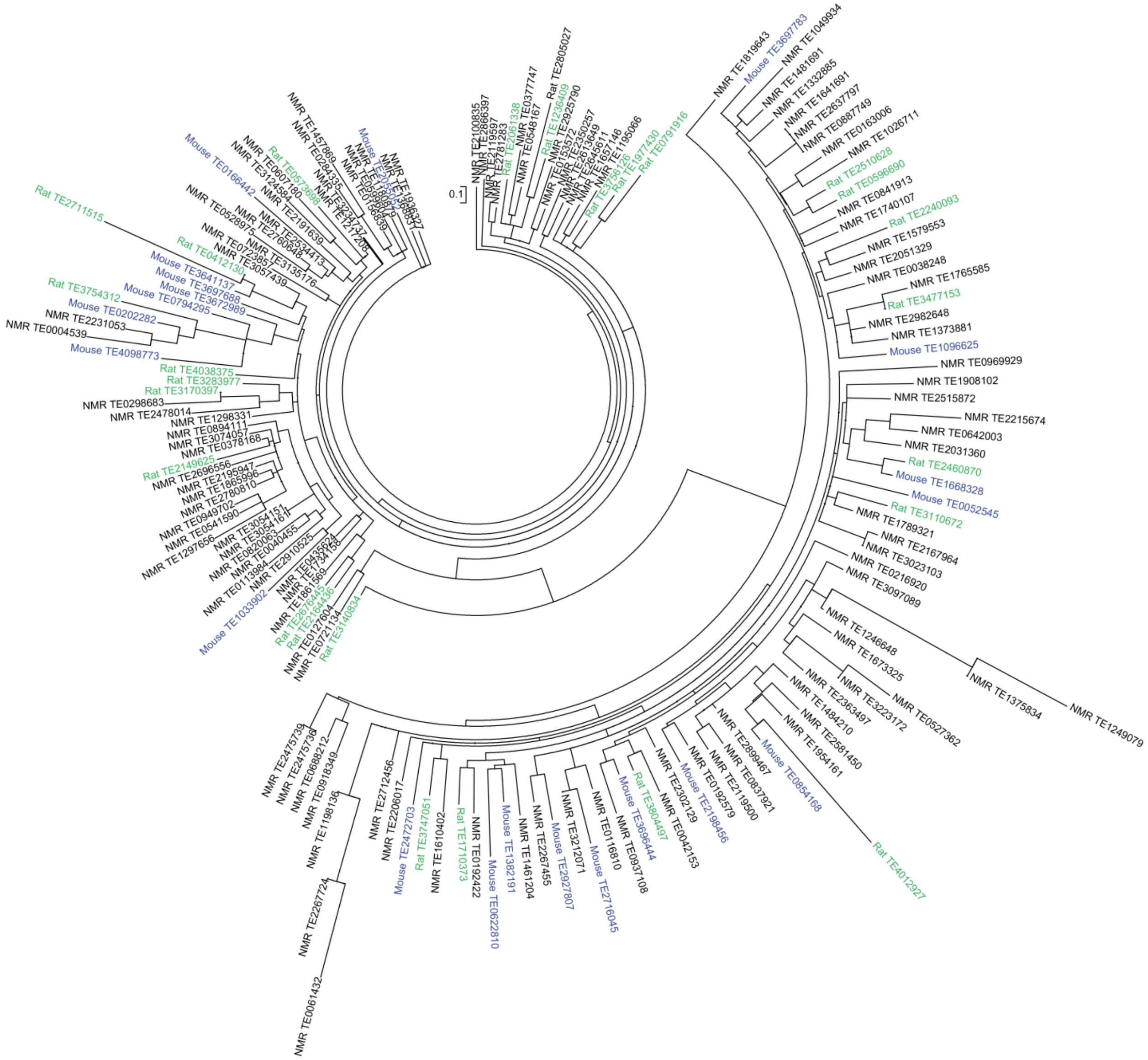
	NMR		Mouse		Rat		Human	
Type	#base	%genome	#base	%genome	#base	%genome	#base	%genome
DNA	57,535,346	2.16	63,664,784	2.34	66,827,052	2.46	102,421,953	3.30
LINE	366,472,637	13.75	495,345,238	18.23	529,530,175	19.48	543,012,030	17.51
LTR	129,574,509	4.86	283,890,666	10.45	220,755,614	8.12	257,192,185	8.29
SINE	118,286,004	4.44	166,730,468	6.14	144,646,502	5.32	349,449,456	11.27
Other	964,138	0.04	7,550,020	0.28	6,774,637	0.25	26,416,214	0.85
Unknown	2,485,954	0.09	43,715,625	1.61	50,716,762	1.87	4,757,709	0.15
Total	666,686,440	25.02	1,024,177,758	37.70	978,028,272	35.97	1,257,671,677	40.55



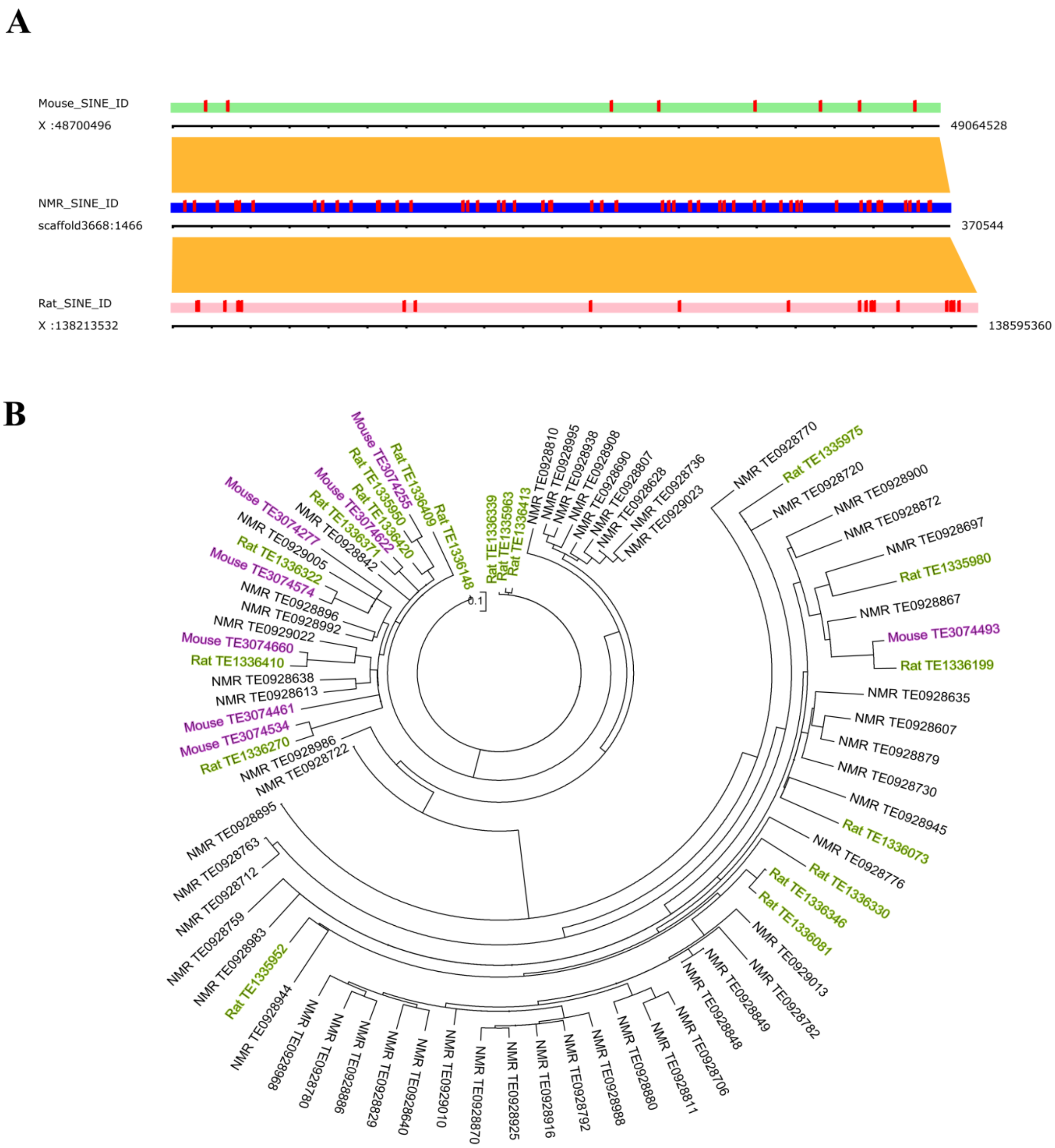
Supplementary Fig. 5. Divergence distribution of classified TE families. To analyze divergence, classified transposal families in NMR, human, mouse and rat genomes were aligned onto the consensus in Repbase.

Supplementary Table 5. TE statistics in four mammalian genomes.

Species		NMR			Rat			Mouse			Human	
TE/Class	**Copy Number**	**#Base**	**% genome**	**Copy Number**	**#Base**	**% genome**	**Copy Number**	**#Base**	**% genome**	**Copy Number**	**#Base**	**% genome**
SINE/ID	471,682	41,069,634	1.541	132,824	12,848,281	0.473	22,334	1,565,639	0.058	562	11,350	0.000
SINE/Alu	415,464	52,951,455	1.987	314,869	35,830,459	1.318	499,693	61,010,293	2.246	1,097,378	302,747,395	11.361
DNA/TcMar	95,324	19,318,104	0.725	68,621	7,936,947	0.292	65,632	8,007,560	0.295	130,815	38,572,160	1.447
SINE/B4	74,198	6,663,303	0.250	284,608	43,467,057	1.599	311,992	47,697,372	1.756	97,447	7,405,030	0.278
DNA/En-Spm	39,195	2,671,612	0.100	185,309	14,057,557	0.517	171,360	13,139,756	0.484	67,973	5,848,967	0.219
DNA/Sola	18,087	1,404,228	0.053	140,749	15,288,002	0.562	119,518	12,213,565	0.450	25,407	2,790,482	0.105
LTR/Gypsy	13,997	1,299,985	0.049	68,499	5,414,606	0.199	57,386	4,570,581	0.168	20,855	2,800,800	0.105
DNA/Maverick	13,208	869,699	0.033	66,515	5,002,471	0.184	59,360	4,417,611	0.163	16,763	1,183,955	0.044
DNA/Harbinger	1,786	111,586	0.004	5,738	399,304	0.015	6,123	419,524	0.015	1,974	134,432	0.005
SINE/B2	83	3,811	0.000	304,507	50,842,159	1.870	339,480	56,267,391	2.071	1	51	0.000



Supplementary Fig 6. Phylogenetic tree of intact ID elements in NMR, mouse and rat genomes. This analysis enabled identification of the copies descended from the ancestor as well as analyses of their gain and loss. The cluster of IDs in NMR suggests their expansion in this organism.



Supplementary Fig 7. Expansion of an ID element in NMR. (A) This case shows ID expansion in NMR in a region of rodent synteny. (B) Phylogenetic tree of IDs in this synteny region collected from NMR, mouse and rat genomes.

3 Gene annotation

3.1 Gene annotation pipeline and evaluation of gene quality

To predict genes in the NMR genome, we used both homology-based and *de novo* methods. In addition, RNA-seq data were incorporated. For the homology-based prediction, human and mouse proteins were downloaded from Ensembl (release 56) and mapped onto the genome using TblastN[5]. Then, homologous genome sequences were aligned against the matching proteins using Genewise[6] to define gene models. For *de novo* prediction, Augustus[7] and Genscan[8] were employed to predict coding genes, using appropriate parameters. RNA-seq data were mapped to genome using Tophat[9], and transcriptome-based gene structures were obtained by cufflinks (http://cufflinks.cbcb.umd.edu/). Finally, homology-based, *de novo* derived and transcript gene sets were merged to form a comprehensive and non-redundant reference gene set using GLEAN (http://sourceforge.net/projects/glean-gene/), removing all genes with sequences less than 50 amino acid as well as those that only had *de novo* support. We obtained a reference gene set that contained 22,561 NMR genes.

Supplementary Table 6. Statistics of predicted protein-coding genes.

Species	*Gene set number*	*Complete ORF*	*%*	*Single exon gene*	*%*	*Average transcript length (bp)*	*Average ORF length (bp)*	*Average exons per gene*	*Average exon length (bp)*	*Average intron length (bp)*
NMR	22,561	19,137	84.82	3,930	17.42	32,533	1,439	8.05	178.73	4,410
Human	22,389	20,098	89.77	3,318	14.82	44,855	1,560	8.96	174.08	5,436
Mouse	23,317	21,196	90.9	4,648	19.93	33,684	1,481	8.37	176.82	4,366
Rat	22,841	16,745	73.31	3,552	15.55	30,892	1,452	8.59	169.06	3,879

3.2 Functional annotation of NMR genes

Functions of NMR genes were assigned based on the best match derived from the alignments to proteins annotated in SwissProt and TrEMBL[10] databases using Blastp. We annotated motifs and domains using InterPro[11] by searching against publicly available databases, including Pfam, PRINTS, PROSITE, ProDom, and SMART. Descriptions of gene products included Gene Ontology[12]; this information was retrieved from InterPro. We also mapped the NMR reference genes to KEGG[13] pathway maps by searching KEGG databases and finding the best hit for each gene.

Supplementary Table 7. Functional classification of NMR genes by various methods.

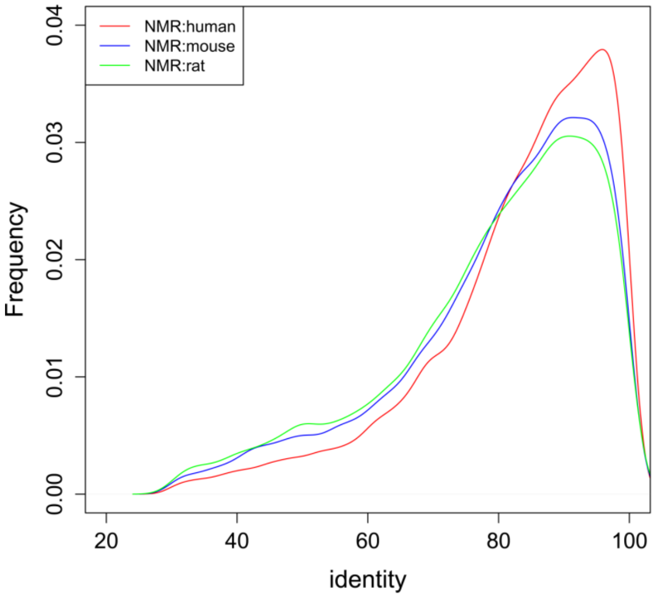
		Number	Percent (%)
Total		22,561	100.00
Annotated	Swissprot	21,922	97.17
	TrEMBL	21,856	96.88
	KEGG	16,917	74.98
	InterPro	18,855	83.57
	GO	14,602	64.72
Unannotated		450	1.99

3.3 Orthology relationship

To determine orthology relationships between NMR and other mammalian proteins, nucleotide and protein data for three mammals (human, mouse and rat) were downloaded from the Ensembl database (release 56). For genes with alternative splicing variants, the longest transcripts were selected to represent the genes. We then subjected human, mouse, rat and NMR proteins to Blastp analysis with the similarity cutoff of $e=1e^{-5}$. With the NMR protein set used as a reference, we found the best hit for each NMR protein in other species, with the criteria that more than 30% of the aligned sequence showed an identity above 30%. Reciprocal best-match pairs were defined as orthologs. Statistics of NMR and other mammalian orthologs is shown in Supplementary Table 8.

Supplementary Table 8. Orthologous relationship between NMR and other mammals.

	Ortholog number
NMR:human	17,030
NMR:mouse	17,525
NMR:rat	17,324



Supplementary Fig. 8. Sequence identity of NMR proteins in comparison with human, mouse and rat proteins. Although NMR and human share fewer orthologs compared to the NMR/mouse and NMR/rat pairs, the identity of NMR/human orthologs is slightly higher.

4 Genome evolution

4.1 Identification of synteny

To detect synteny blocks between NMR and other mammals, pairwise whole-genome alignment was performed using LASTZ with parameters T=2, and Y=9400 (http://www.bx.psu.edu/miller_lab/). ChainNet, which can accommodate inversions, translocations, duplications, large-scale deletions, and overlapping deletions, was used to combine traditional alignments into larger structures.

Supplementary Table 9. Syntenic regions between NMR and other mammalian genomes.

Organism	Size (G)	Synteny(G)	%	Query	Size (G)	Synteny (G)	%	# Blocks
Dog	2.53	2.24	88.59	NMR	2.66	2.41	90.53	18,124
Human	3.1	2.74	88.22	NMR	2.66	2.46	92.31	24,999
Mouse	2.72	2.16	79.38	NMR	2.66	2.22	83.28	41,225
Rat	2.72	2.16	79.49	NMR	2.66	2.12	79.65	41,779

4.2 Lineage-specific indels

Following synteny analysis, MULTIZ[13] was used to integrate all pairwise alignments together to get the Conserved Elements among human, mouse, rat and NMR genomes. For blocks longer than 1,000 bp we counted species-specific short indels according to the align data; the indels located within 50 bp of the end of the block and the pairs of indels with the distance less than 50 bp were filtered out.

Supplementary Table 10. Statistics of lineage-specific genome rearrangements.

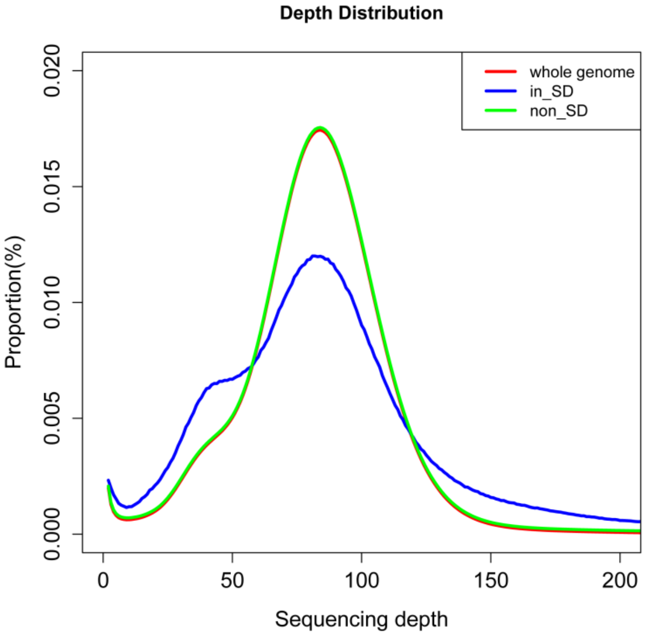
Organism	InDel/Myr	Ins/Myr	Del/Myr
NMR	51.61	16.02	35.60
Mouse	70.65	38.05	32.60
Rat	90.20	44.55	45.65
Human	33.48	14.90	18.58

4.3 Segmental duplications

We used whole-genome assembly comparison to identify segmental duplications (SDs)[15]. The self-alignment for each genome was implemented using LASTZ with parameters T=2 and Y=9400. SDs were defined as two sequences larger than 1 kb with identity higher than 90%, but lower than 98%, to exclude improperly assembled allelic variants due to the draft status of the genome. For comparison, the same pipeline and parameters were applied to human (hg18), mouse (mm8) and rat (rn4) genomes. The SD analysis revealed that NMR had the lowest proportion of SD compared to the other three mammals as shown in Supplementary Table 11. To check quality of detected SD in the NMR genome, we compared the depth distribution of SD and non-SD regions. If the SDs were due to false positives, i.e., more copies detected in the assembly than actually present, the depth of such regions calculated based on short read alignment should have been lower than in the whole genome. In addition, if some SD copies in the NMR genome were not assembled, one would expect that the reads would not be aligned to such regions if they are not present in the assembly (completely missing) or if they show a higher depth (some copies are missing). The depth of SD regions showed a similar pattern (except for two fat tails) to non-SD regions, suggesting no large-scale problems with our detected SDs.

Supplementary Table 11. Statistics of segmental duplications in NMR and other mammalian genomes.

Organism	Genome size	SD size	%
NMR	2,664,766,285	85,278,445	3.20
Human	3,101,788,170	111,338,389	3.59
Mouse	2,644,077,689	124,351,298	4.70
Rat	2,718,897,321	89,752,747	3.30



Supplementary Fig. 9. Depth distribution of SD regions. The depth density of SD regions, non-SD regions and the whole genome was calculated and plotted.

5 Gene evolution

5.1 Gene family clusters

DNA and protein data for three mammals (human, mouse and rat) were downloaded from the Ensembl database (release 56). For genes with alternative splicing variants, the longest transcripts were selected to represent the genes. We used Treefam methodology[17] to define a gene family as a group of genes that descended from a single gene in the last common ancestor of considered species.

1) Blastp was applied to all protein sequences against a database containing a protein dataset of all species with the e-value of 1e-7 and conjoined fragmental alignments for each gene pair by Solar. We assigned a connection (edge) between the two nodes (genes) if more than 1/3 of the region aligned to both genes. An Hscore that ranged from 0 to 100 was used to weigh the similarity (edge). For two genes, G1 and G2, the Hscore was defined as a score (G1G2)/max (score (G1G1), score (G2G2)) (the score here is the raw Blast score).

2) Extraction of gene families (clustering by Hcluster_sg). We used the average distance for the hierarchical clustering algorithm, requiring the minimum edge weight (Hscore) to be larger than 5, and the minimum edge density (total number of edges/theoretical number of edges) to be larger than 1/3.

5.2 Phylogenetic tree and divergence time

We constructed a phylogenetic tree of NMR and several other sequenced mammals (dog, human, rhesus macaque, rabbit, mouse and rat) using single-copy orthologous genes. 4-fold degenerate sites were extracted from each family and concatenated to one supergene for one species. Modeltest[18] was used to select the best substitution model and Mrbayes[19] to reconstruct the phylogenetic tree. The chain length was set to 50,000,000 (1 sample/1000 generations) and the first 1,000 samples were burned in. The transition/transversion ratio was estimated as a free parameter. Other parameters were set with the default setting. The BRMC approach was used to estimate the species divergence time using the program MULTIDIVTIME[20,21], which was implemented using the Thornian Time Traveller

(T3) package (ftp://abacus.gene.ucl.ac.uk/pub/T3/).

5.3 Analyses of gene gain and loss

Orthology information was obtained as described above. Since it showed synteny information at the protein level, it could be used to analyze gene gain and loss between human and NMR. In the protein synteny blocks, if a human protein had no NMR ortholog, and excluding false positive predictions that could be caused by annotation or genome assembly (gap > 5%), this protein could be defined as either being lost in the NMR lineage or gained in the human lineage. Using NMR as a reference to generate the orthology relationship, we applied this procedure to identify genes gained in the NMR lineage compared to the human lineage.

Supplementary Table 14. Gene gain/loss in NMR in comparison with other mammals.

	Human	Mouse	Rat
NMR gain	750	739	414
NMR gain/Myr	8.0559	10.109	5.6635
NMR loss	320	448	246
NMR loss/Myr	3.4372	6.1286	3.3653

Supplementary Table 15. GO enrichment of genes that were lost in NMR.

GO_ID	GO_Term	GO_Class	Adjusted p-value
GO:0030529	ribonucleoprotein complex	CC	0.023655
GO:0003735	structural constituent of ribosome	MF	0.023655
GO:0005840	ribosome	CC	0.023655
GO:0004550	nucleoside diphosphate kinase activity	MF	0.023655
GO:0006183	GTP biosynthetic process	BP	0.023655
GO:0006228	UTP biosynthetic process	BP	0.023655
GO:0006241	CTP biosynthetic process	BP	0.023655
GO:0006412	translation	BP	0.046916

5.4 Pseudogene identification and selective constraints of NMR pseudogenes

We used human proteins to call homologs in the NMR genome located in synteny blocks (synteny blocks were determined by the human/NMR whole genome alignment). For frameshift and premature termination events occurring in homologous regions, we manually examined genomic and transcriptomic read mapping quality of frameshift and premature termination loci. Cases with high mapping quality, excluding any SNPs or indels, were inferred as mutations, which in turn identified pseudogenes. To examine selective constraints on NMR pseudogenes, we estimated the rate ratio (ω) of nonsynonymous to synonymous substitutions using PAML. We aligned the NMR pseudogene sequences with their human, mouse and rat homologs using the program Muscle. We then compared a series of evolutionary models in the likelihood framework using the species tree of human, mouse, rat, and NMR. The branch model was used to detect the average ω across the tree ($\omega0$), the ω of the NMR branch ($\omega2$) and the ω of all other branches ($\omega1$). Then, the chi-square test was used to test whether $\omega2$ is significantly higher than $\omega1$ and $\omega0$, with inference that these genes escaped the selection constrain after becoming pseudogenes.

5.5 Positively selected genes

To detect genes that evolved under positive selection, we used PAML, a Maximum-Likelihood method of molecular evolution[22,23]. Specifically, we used the PAML's branch-site test of positive selection[24,25] to test for positive selection along the NMR branch. We compared ModelA1, in which sites may evolve neutrally and under purifying selection with ModelA that allows sites to be also under positive selection. P-values were computed using the X^2 statistic adjusted by the fdr method to allow for multiple testing. Alignment quality is of major importance for studies of positive selection as alignment errors can lead to unacceptable high false positives using the branch-site model[26]. We used PRANK[27] which differs from other alignment tools in that it utilizes evolutionary information in determining where to place a gap. Studies on the branch-site test and on other PAML models support PRANK to be the alignment tool of choice[26,28]. We filtered the PRANK alignments by gblocks[29,30] and excluded genes with sequence properties that often lead to false positives, such as genes with high proportion of low complexity or disordered regions, ubiquitous domains, repeats, and transmembrane and coiled-coil regions.

Supplementary Table 18. Positively selected genes.

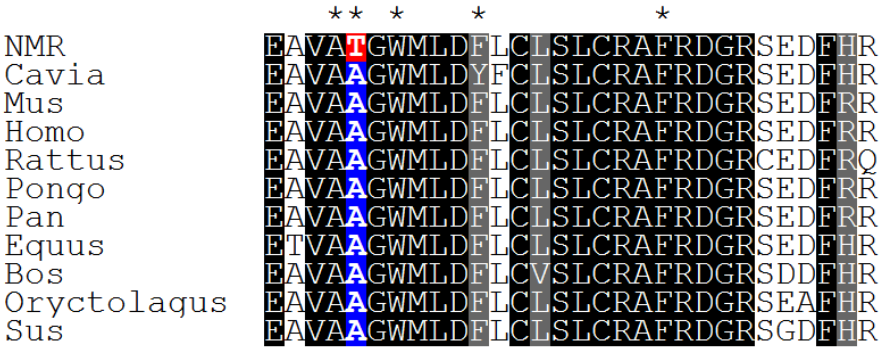
Gene Symbol	**Protein ID**	***H. glaber* protein**	**FDR**	**Gene description**
COL4A2	P08572	HGL_H00000378340	0.0001	Collagen alpha-2(IV) chain
CCDC162	A2VCL2	HGL_H00000402649	0.0000	Coiled-coil domain-containing protein 162
PCDHA3	Q9Y5H8	HGL_H00000367372	0.0002	Protocadherin alpha-3
RHOBTB2	Q9BYZ6	HGL_H00000251822	**0.0002**	**Rho-related BTB domain-containing protein 2**
ROBO4	Q8WZ75	HGL_H00000304945	**0.0002**	**Roundabout homolog 4**
PEAR1	Q5VY43	HGL_H00000344465	0.0002	Platelet endothelial aggregation receptor 1
C1orf173	Q5RHP9	HGL_H00000322609	0.0003	Uncharacterized protein C1orf173
TMPO	P42167	HGL_H00000266732	0.0003	Lamina-associated polypeptide 2
ZNF167	Q9P0L1	HGL_H00000273320-1	0.0003	Zinc finger protein 167
FLG2	Q5D862	HGL_H00000357789	0.0003	Filaggrin-2
ABCA9	Q8IUA7	HGL_H00000411772	0.0003	ATP-binding cassette subfamily A member 9
CARD6	**Q9BX69**	**HGL_N10017264**	**0.0003**	**Caspase recruitment domain-containing protein 6**
MEGF6	O75095	HGL_H00000398045-1	0.0005	Multiple epidermal growth factor-like domains protein 6
CCDC15	Q0P6D6	HGL_H00000341684	0.0009	Coiled-coil domain-containing protein 15
FGFR2	P21802	HGL_H00000309878	0.0011	Fibroblast growth factor receptor 2
C12orf43	Q96C57	HGL_H00000288757-1	0.0013	Uncharacterized protein C12orf43
PCDHAC2	Q9Y5I4	HGL_H00000377862-2	0.0013	Protocadherin alpha-C2
C12orf43	Q5VWT5	HGL_H00000345972	0.0015	Uncharacterized protein C1orf168
DPEP1	**P16444**	**HGL_H00000261615**	**0.0015**	**Dipeptidase 1**
TAAR2	Q9P1P5	HGL_H00000275216	0.0015	Trace amine-associated receptor 2
PCDHGB1	Q9Y5G3	HGL_H00000367345-2	0.0015	Protocadherin gamma-B1
C2orf71	A6NGG8	HGL_H00000332809	0.0021	Uncharacterized protein C2orf71
SLC9A11	**Q5TAH2**	**HGL_H00000356687**	**0.0021**	**Sodium/hydrogen exchanger 11**
HIVEP2	P31629	HGL_H00000360069	0.0021	Transcription factor HIVEP2
TBR1	Q16650	HGL_H00000374205	0.0023	T-box brain protein 1
BTF3	P20290	HGL_H00000369965-4	0.0026	Transcription factor BTF3
NCKAP5L	**Q9HCH0**	**HGL_H00000387128**	**0.0031**	**Nck-associated protein 5-like**
KIAA0319	Q5VV43	HGL_H00000367459	0.0031	Dyslexia-associated protein
PAK7	**Q9P286**	**HGL_H00000367679**	**0.0047**	**Serine/threonine-protein kinase PAK 7**
ZNRD1-AS1	Q2KJ03	HGL_M00000048695-2	0.0047	Putative uncharacterized protein
DNAJC1	Q96KC8	HGL_H00000366179	**0.0062**	**DnaJ homolog subfamily C member 1**
TEP1	**Q99973**	**HGL_H00000262715**	**0.0067**	**Telomerase protein component 1**
SLC19A3	Q9BZV2	HGL_H00000258403-3	0.0071	Thiamine transporter 2
ABCC10	Q5T3U5	HGL_H00000361608	0.0076	Multidrug resistance-associated protein 7

OR56A3	Q8NH54	HGL_H00000331572-1	0.0076	Olfactory receptor 56A3
RPRD1A	Q96P16	HGL_H00000349955-2	**0.0076**	**Regulation of nuclear pre-mRNA domain-containing protein 1A**
COL24A1	Q17RW2	HGL_H00000359603	0.0076	Collagen alpha-1(XXIV) chain
KCNQ1	P51787	HGL_N10021971	0.0076	Potassium voltage-gated channel subfamily KQT member 1
COL3A1	P02461	HGL_H00000304408	0.0080	Collagen alpha-1(III) chain
MYL6	P60660	HGL_H00000293422	0.0083	Myosin light polypeptide 6
DMRTA2	Q96SC8	HGL_H00000360500	0.0083	Doublesex- and mab-3-related transcription factor A2
E2F4	Q16254	HGL_H00000368686	**0.0083**	**Transcription factor E2F4**
OLFM4	Q6UX06	HGL_H00000219022-1	**0.0086**	**Olfactomedin-4**
CCDC27	Q2M243	HGL_H00000294600	0.0086	Coiled-coil domain-containing protein 27
GPR112	Q8IZF6	HGL_H00000359686	0.0096	Probable G-protein coupled receptor 112
DHRS11	Q6UWP2	HGL_H00000251312	0.0101	Dehydrogenase/reductase SDR family member 11
GTF2F2	P13984	HGL_H00000340823	0.0105	General transcription factor IIF subunit 2
MTMR2	Q13614	HGL_H00000345752-1	0.0105	Myotubularin-related protein 2
HIPK1	Q86Z02	HGL_H00000407442-1	0.0105	Homeodomain-interacting protein kinase 1
LPLUNC1	Q8TDL5	HGL_H00000253354	0.0108	Long palate, lung and nasal epithelium carcinoma-associated protein 1
COL4A2	P08572	HGL_H00000353654	0.0108	Collagen alpha-2(IV) chain
PRAMEF23	A6NMV5	HGL_H00000365363	0.0108	PRAME family member 23
ANKRD26	Q9UPS8	HGL_H00000405112-1	0.0108	Ankyrin repeat domain-containing protein 26
APOBR	Q0VD83	HGL_M00000042028	0.0108	Apolipoprotein B receptor
LGALS8	O00214	HGL_N10000412	0.0108	Galectin-8
FOLR1	P15328	HGL_H00000377284	0.0116	Folate receptor alpha
SLAMF7	Q9NQ25	HGL_H00000263285	0.0117	SLAM family member 7
SCARF1	Q14162	HGL_H00000263071	0.0122	Scavenger receptor class F member 1
LMAN1	P49257	HGL_H00000251047	0.0132	Protein ERGIC-53
PRAMEF12	O95522	HGL_H00000350358-1	0.0132	PRAME family member 12
SPTB	P11277	HGL_H00000374373	0.0132	Spectrin beta chain, erythrocyte
ADAMTS7	Q9UKP4	HGL_H00000258883	0.0164	Metalloprotease
GAL	P22466	HGL_N10009473	0.0178	Galanin
HEG1	Q9ULI3	HGL_H00000311502	0.0185	Protein HEG homolog 1
COL18A1	P39060	HGL_H00000352798	0.0187	Collagen alpha-1(XVIII) chain
MAGEA10	P43363	HGL_H00000244096-9	0.0197	Melanoma-associated antigen 10
FCRL1	Q96LA6	HGL_H00000292389	0.0209	Fc receptor-like protein 1
ZNF167	Q9P0L1	HGL_H00000415358-2	0.0221	Zinc finger protein 167
GABRQ	Q9UN88	HGL_H00000359329	0.0237	Gamma-aminobutyric acid receptor subunit theta
IGBP1	P78318	HGL_H00000363661-3	0.0237	Immunoglobulin-binding protein 1
PRKD2	Q9BZL6	HGL_H00000408285-1	0.0237	Serine/threonine-protein kinase D2
LETM1	O95202	HGL_H00000305653-2	0.0240	LETM1 and EF-hand domain-containing protein 1, mitochondrial

DPP3	Q9NY33	HGL_H00000353701	0.0240	Dipeptidyl peptidase 3
MAGEA10	P43363	HGL_H00000349085	0.0241	Melanoma-associated antigen 10
COL23A1	Q86Y22	HGL_H00000375069	0.0242	Collagen alpha-1(XXIII) chain
PRSS58	Q8IYP2	HGL_H00000414461-2	0.0243	Serine protease 58
PARP14	Q460N5	HGL_H00000418194	0.0251	Poly [ADP-ribose] polymerase 14
ANO1	Q5XXA6	HGL_H00000347454	0.0270	Anoctamin-1
PLCZ1	Q86YW0	HGL_H00000266505	0.0271	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase zeta-1
WDSUB1	Q8N9V3	HGL_H00000380377	0.0271	WD repeat, SAM and U-box domain-containing protein 1
LCN9	Q8WX39	HGL_H00000277526-2	0.0273	Epididymal-specific lipocalin-9
HMX1	Q9NP08	HGL_H00000350549	0.0273	Homeobox protein HMX1
FAM38A	Q92508	HGL_H00000301015	0.0275	Protein PIEZO1
SUGT1	Q9Y2Z0	HGL_H00000367208-1	0.0277	Suppressor of G2 allele of SKP1 homolog
AGRP	O00253	HGL_H00000290953	0.0280	Agouti-related protein
DNM1L	O00429	HGL_H00000244426-1	0.0288	Dynamin-1-like protein
GATA1	P15976	HGL_H00000398566	0.0288	Erythroid transcription factor
CLDN8	P56748	HGL_H00000286809	0.0293	Claudin-8
NTN5	Q8WTR8	HGL_H00000270235	0.0304	Netrin-5
RBM4	Q9BWF3	HGL_H00000309166-1	0.0304	RNA-binding protein 4
C16orf86	Q6ZW13	HGL_H00000384117	0.0304	Uncharacterized protein C16orf86
STRC	Q7RTU9	HGL_H00000401513-2	0.0309	Stereocilin
PPP1R3F	Q6ZSY5	HGL_H00000055335	0.0309	Protein phosphatase 1 regulatory subunit 3F
SPI1	P17947	HGL_H00000227163	0.0309	Transcription factor PU.1
TERF1	P54274	HGL_H00000276603-1	0.0309	Telomeric repeat-binding factor 1
CEL	P19835	HGL_H00000361151-2	0.0309	Bile salt-activated lipase
ANGPT4	Q9Y264	HGL_H00000371347-3	0.0309	Angiopoietin-4
VPS13A	Q96RL7	HGL_H00000365834	0.0312	Vacuolar protein sorting-associated protein 13A
C19orf21	Q8IVT2	HGL_H00000215582	0.0328	Uncharacterized protein C19orf21
SLC30A5	Q8TAD4	HGL_H00000379836-2	0.0354	Zinc transporter 5
SERPINA1	P01009	HGL_H00000416066-2	0.0354	Alpha-1-antitrypsin
SYCP2	Q9BX26	HGL_H00000350162	0.0358	Synaptonemal complex protein 2
MEIR5	Q86XK3	HGL_H00000338089-2	0.0360	Swi5-dependent recombination DNA repair protein 1 homolog
GREB1L	Q9C091	HGL_H00000412060	0.0362	GREB1-like protein
P4HA	P13674	HGL_H00000307318	0.0386	Prolyl 4-hydroxylase subunit alpha-1
C17orf66	A2RTY3	HGL_H00000309560	0.0386	Uncharacterized protein C17orf66
RRP1B	Q14684	HGL_H00000339145	0.0386	Ribosomal RNA processing protein 1 homolog B
C14orf43	Q6PJG2	HGL_H00000377634	0.0386	Uncharacterized protein C14orf43
RBM28	Q9NW13	HGL_N10014876	0.0386	RNA-binding protein 28
MLL5	Q8IZD2	HGL_H00000257745	0.0390	Histone-lysine N-methyltransferase MLL5
COL7A1	Q02388	HGL_H00000332371	0.0390	Collagen alpha-1(VII) chain

NUMB	P49757	HGL_H00000347169-1	0.0114	Protein numb homolog
GSTO1	P78417	HGL_H00000358727	0.0395	Glutathione S-transferase omega-1
MAGEA10	P43363	HGL_H00000244096-6	0.0395	Melanoma-associated antigen 10
SPINK5	Q9NQ38	HGL_H00000352936	0.0395	Serine protease inhibitor Kazal-type 5
PDCD5	O14737	HGL_H00000388543-4	0.0413	Programmed cell death protein 5
CD320	Q9NPF0	HGL_M00000005352	0.0416	CD320 antigen
KRT31	Q15323	HGL_H00000377572	0.0419	Keratin, type I cuticular Ha1
CD34	P28906	HGL_H00000310036	0.0419	Hematopoietic progenitor cell antigen CD34
FAM65C	Q96MK2	HGL_H00000332663	0.0435	Protein FAM65C
PECR	Q9BY49	HGL_H00000265322-2	0.0443	Peroxisomal trans-2-enoyl-CoA reductase
KIAA1468	Q9P260	HGL_H00000256858	0.0444	LisH domain and HEAT repeat-containing protein KIAA1468
PRRT2	Q7Z6L0	HGL_H00000351608	0.0444	Proline-rich transmembrane protein 2
HIRIP3	Q9BW71	HGL_H00000279392	0.0453	HIRA-interacting protein 3
AVP	P01185	HGL_H00000369647-2	0.0453	Vasopressin-neurophysin 2-copeptin
MYBBP1A	Q9BQG0	HGL_H00000370968-2	0.0453	Myb-binding protein 1A
AIRE	O43918	HGL_H00000291582	0.0453	Autoimmune regulator
ZNF592	Q92610	HGL_H00000299927	0.0453	Zinc finger protein 592
CCL8	P80075	HGL_H00000378118-2	0.0453	C-C motif chemokine 8
MRPL28	Q13084	HGL_H00000380843	0.0453	Melanoma-associated antigen recognized by T lymphocytes
FAM195A	Q9BUT9	HGL_H00000305138-2	0.0453	Protein FAM195A
ATF6	P18850	HGL_N10009489	0.0480	Cyclic AMP-dependent transcription factor ATF-6 alpha
GAS2L2	Q8NHY3	HGL_H00000254466	0.0480	GAS2-like protein 2
HDC	P19113	HGL_H00000267845	0.0480	Histidine decarboxylase
AP2A2	O94973	HGL_H00000327694	0.0480	AP-2 complex subunit alpha-2
ADAMTS13	Q76LX8	HGL_H00000360997	0.0480	A disintegrin and metalloproteinase with thrombospondin motifs 13
CACNA1C	Q13936	HGL_H00000385724	0.0480	Voltage-dependent L-type calcium channel subunit alpha-1C
TMEM31	Q5JXX7	HGL_H00000316940	0.0485	Transmembrane protein 31
CEACAM16	Q2WEN9	HGL_H00000379974	0.0485	Carcinoembryonic antigen-related cell adhesion molecule 16
BDP1	A6H8Y1	HGL_H00000351575	0.0497	Transcription factor TFIIIB component B" homolog
KIAA1009	Q5TB80	HGL_H00000385215	0.0497	Protein QN1 homolog

141 genes were identified by PAML’s branch-site test of positive selection. Among the first 45 genes (with FDR<0.01), the genes shown in bold were checked manually. Some of the genes in this table, especially those not shown in bold, may be false-positives. Certain protein properties may increase the chance of misalignment, misannotation and positive selection. We considered proteins with >25% coiled coil domains, low complexity and/or disordered regions as problematic. Also, proteins with overlapping domains causing multiple representations, uncharacterized proteins, collagens, Zn-finger proteins, olfactory receptors and other large families or clustered arrangements were considered problematic. We used SMART to examine protein properties.



```
                 ** *   *        *
NMR          EAVATGWMLDFLCLSLCRAFRDGRSEDFHR
Cavia        EAVAAGWMLDYFCLSLCRAFRDGRSEDFHR
Mus          EAVAAGWMLDFLCLSLCRAFRDGRSEDFRR
Homo         EAVAAGWMLDFLCLSLCRAFRDGRSEDFRR
Rattus       EAVAAGWMLDFLCLSLCRAFRDGRCEDFRQ
Pongo        EAVAAGWMLDFLCLSLCRAFRDGRSEDFRR
Pan          EAVAAGWMLDFLCLSLCRAFRDGRSEDFRR
Equus        ETVAAGWMLDFLCLSLCRAFRDGRSEDFHR
Bos          EAVAAGWMLDFLCVSLCRAFRDGRSDDFHR
Oryctolagus  EAVAAGWMLDFLCLSLCRAFRDGRSEAFHR
Sus          EAVAAGWMLDFLCLSLCRAFRDGRSGDFHR
```

Supplementary Fig. 10. NMR-specific amino acid change in *TERF*. Asterisks indicate residues involved in telomere binding in human TRF1 protein encoded by *TERF1*. The Ala75Pro mutation in the human protein is known to inhibit dimerization of TRF1 and telomere binding. The same amino acid changed in the NMR sequence.

5.6 Identification of NMR proteins with unique amino acid changes

NMR proteins were aligned to UCSC multiple protein alignments through pairwise alignment with human proteins. Hits with less than 25% identity were dismissed, and coordinates of all amino acid differences between human and NMR proteins were stored for further analysis. The conservation of amino acids in the corresponding positions within the multi-way UCSC Vertebrate Alignment was examined. A total of 42,399 candidates were considered. The following organisms were included in the analysis: *Homo sapiens, Pan troglodytes*, *Pongo pygmaeus abelii*, *Macaca mulatta*, *Papio hamadryas*, *Callithrix jacchus*, *Tarsier syrichta*, *Microcebus murinus*, *Otolemur garnettii*, *Tupaia belangeri*, *Mus musculus*, *Rattus norvegicus*, *Dipodomys ordii*, *Cavia porcellus*, *Spermophilus tridecemlineatus*, *Oryctolagus cuniculus*, *Ochotona princeps*, *Vicugna pacos*, *Bos taurus*, *Equus caballus*, *Felis catus*, *Canis lupus familiaris*, *Myotis lucifugus*, *Pteropus vampyrus*, *Erinaceus europaeus*, *Sorex araneus*, *Loxodonta africana*, *Echinops telfairi*, *Macropus eugenii*, *Monodelphis domestica*, *Ornithorhynchus anatinus, Xenopus tropicalis, Tetraodon nigroviridis, Takifugu rubripes*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Danio rerio*, and *Petromyzon marinus*. We selected sequences containing amino acids that are conserved in all available sequences in the 36 genomes, except for NMR. To remove redundant hits, an additional Blastp analysis was carried out that applied bidirectional best hit criteria. The remaining 95 candidates were analyzed for conservation of the region within which amino acid changes occurred. Finally, transcriptome data were utilized to verify the *H. glaber* genes and exclude gene misprediction and misannotation events. This analysis yielded 39 vertebrate proteins that uniquely changed one or more conserved amino acids in the NMR lineage.

Supplementary Table 19. Conserved proteins that uniquely changed amino acids in NMR.

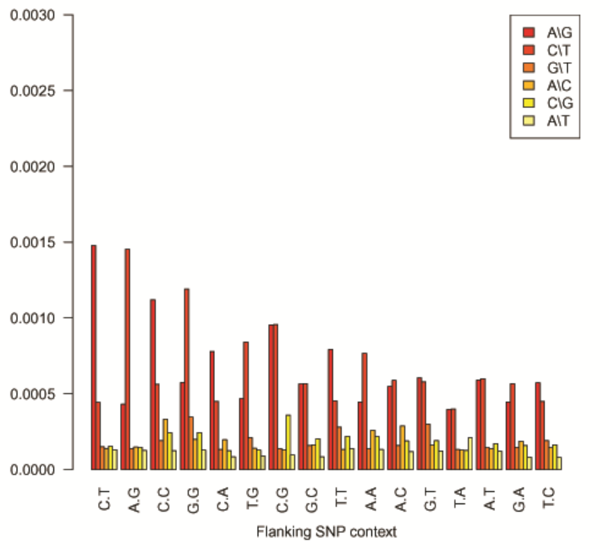
Gene name	**Symbol**	***H. glaber* protein**	**Amino acid change**
A disintegrin and metalloproteinase with thrombospondin motifs 1	*ADAMTSL1*	HGL_H00000369921	V404E
Aldolase B	*ALDOB*	HGL_H00000363988	N169V
APEX nuclease	*APEX1**	HGL_H00000381111	E40D
Bardet-Biedl syndrome 7	*BBS7*	HGL_H00000264499	D412N
Cadherin-20	*CDH20**	HGL_H00000262717	D547N
Chloride intracellular channel protein 6	*CLIC6*	HGL_H00000353959	P592L
Collagen alpha-2(V) chain	*COL5A2*	HGL_H00000364000	I994V, P1012Q
Cysteinyl leukotriene receptor 2	*KPG_011*	HGL_H00000282018	Y123C, R136H
Dedicator of cytokinesis 5	*DOCK5*	HGL_H00000276440	E391D
DNA topoisomerase 2-alpha	*TOP2A**	HGL_H00000269577	N1126A
E3 ubiquitin-protein ligase HERC2	*HERC2*	HGL_H00000261609	L3893S
Exonuclease 3'-5' domain-like protein 2	*B4DIH6*	HGL_H00000387331	D410G, Y429C
FRAS1-related extracellular matrix protein 2	*FREM2*	HGL_H00000280481	R1128S
G1/S-specific cyclin-E1	*CCNE1**	HGL_H00000262643	A335V
Hemicentin-1	*HMCN1*	HGL_H00000271588	V4075I
Hypothetical protein LOC79624	*UPF0364*	HGL_H00000356263	G375A
Inactive ubiquitin C-terminal hydrolase 54	*USP54*	HGL_H00000345216	I151V
Integrator complex subunit 9	*INTS9*	HGL_H00000398208	A208V
Maternal embryonic leucine zipper kinase	*MELK**	HGL_H00000298048	L133I
Membrane-associated guanylate kinase	*MAGI1**	HGL_H00000385450	R1020H
Mitochondrial uncoupling protein 1	*UCP1*	HGL_H00000262999	G263R
Neuron-derived orphan receptor 1	*NOR-1*	HGL_H00000333122	D441E
Nuclear receptor subfamily 2 group C member 1	*NR2C1*	HGL_H00000333275	Q415H
Oxysterol-binding protein 1	*OSBP*	HGL_H00000263847	M446I, L498I, S521N
Probable G-protein coupled receptor 176	*GPR176*	HGL_H00000299092	Q378E
Probable phospholipid-transporting ATPase IC	*ATP8B1*	HGL_H00000283684	I393L
Protein unc-13 homolog B	*UNC13B*	HGL_H00000380006	L427F
Pumilio homolog 1	*PUM1*	HGL_H00000362846	N773S
Replication factor C	*RFC1**	HGL_H00000371321	K759R
Sodium/glucose cotransporter 4	*SLC5A9*	HGL_H00000236495	S312C
Solute carrier family 25 member 36	*SLC25A36*	HGL_H00000391521-7	G175R
Solute carrier family 30 zinc transporter, member 9	*ZnT-9*	HGL_H00000264451-1	G412S
StAR-related lipid transfer protein 13	*STARD13*	HGL_H00000338785	G673R
Tryptophanyl-tRNA synthetase, cytoplasmic	*WARS*	HGL_H00000347495	S292C
Tubulin epsilon chain	*TUBE1*	HGL_H00000357651	Q321L
Ubiquitin-like modifier-activating enzyme	*ATG7*	HGL_H00000346437	K48R
UPF0505 protein C16orf62	*C16orf62*	HGL_H00000251143	A527G
Vacuolar protein sorting 41 homolog	*VPS41*	HGL_H00000309457	V282A
γ-crystallin	*CRYGS**	HGL_H00000312099	V42A, G45A

* These genes have been designated as cancer genes[31].

5.7 Analysis of genetic variation

We used GATK software to call heterozygous SNP positions within the NMR genome. Overall per nucleotide heterozygosity for NMR is 0.0007. We compared it to the human polymorphism data recently released by Complete Genomics and found that per nucleotide heterozygosity in NMR is lower than in human individuals from sub-Saharan Africa, but comparable to that of human out of Africa populations. In protein coding regions, non-synonymous SNPs were more common than synonymous SNPs (ratio 1.16). This ratio is higher than in humans and much higher than in rodent species.

Although the NMR transition-transversion ratio is very similar to that of other mammalian genomes, the fraction of SNPs with one of the alleles within a hypermutable CpG context was lower than in humans. There were a total of 463,100 such SNPs in NMR out of 1,982,148 SNPs, i.e. 0.23 of the total SNPs were within CpGs, compared to 0.295 in humans (YRI, sample ID NA19238). Mainly this reflects lower CpG density in the NMR genome. The fraction of CpGs is 0.19 of the expected given the GC content. In comparison, this fraction is 0.29 for panda, 0.26 for dog, 0.24 for human, and 0.19 for mouse. However, even though the NMR CpG density is highly similar to that in the mouse genome, 68 Mb were covered by 200 nt windows with the GC content exceeding 0.5 and the CpG density higher than 0.6 of the expected. Only 38 Mb of the sequence were covered by such windows in the mouse genome. Thus, a higher fraction of CpGs reside in CpG islands in the NMR genome. Although genetic variation in the natural mouse population has not be fully characterized, we hypothesize that impact of CpG hypermutability on variation rate in NMR may be lower than in other mammals, including rodents, because CpG di-nucleotides in CpG islands display lower levels of genetic variation.



Supplementary Fig. 11. Context dependency of SNPs in the NMR genome.

6 Transcriptome analyses

6.1 Transcriptome sequencing

Total RNA was isolated from brain, kidney, and liver of newborn, 4-year old, and 20-year old female NMRs as well as from a 4-year old female NMR maintained in a 8% oxygen environment for one week. RNA sequencing libraries were constructed using the Illumina mRNA-Seq Prep Kit. Briefly, oligo(dT) magnetic beads were used to purify polyA containing mRNA molecules. The mRNA was further fragmented and randomly primed during the first strand synthesis by reverse transcription. This procedure was followed by second-strand synthesis with DNA polymerase I to create double-stranded cDNA fragments. The double stranded cDNA was subjected to end repair by Klenow and T4 DNA polymerases and A-tailed by Klenow lacking exonuclease activity. Ligation to Illumina Paired-End Sequencing adapters, size selection by gel electrophoresis and then PCR amplification completed library preparation. The 200 bp paired-end libraries were sequenced using Illumina HiSeq 2000 (90 bp at each end).

Supplementary Table 20. Transcriptome sequencing data statistics.

	Total reads (M)	Total base (G)	Map reads (M)	Reads (%)	Map base (G)	Base (%)	Genome coverage (%)
New-brain	55.1	4.96	47.8	86.8	4.06	81.9	3.96
New-kidney	48.2	4.34	42.5	88.2	3.63	83.6	4.38
New-liver	53.3	4.8	45.7	85.7	3.85	80.2	3.22
4-brain	53.4	4.81	43.7	81.8	3.64	75.7	3.19
4-kidney	50.4	4.54	40.5	80.4	3.35	74	2.91
4-liver	54.5	4.91	45.2	83	3.76	77	2.68
20-brain	58.4	5.25	48.2	82.5	4.05	77.1	3.89
20-liver	52.8	4.75	44.9	85	3.78	80	3.11
20-kidney	56	5	45.4	81.7	3.8	76	3.2
Low-liver	66.7	6	55.67	83.5	4.63	77.2	2.41
Low-kidney	65.8	5.93	52.09	79.1	4.33	73.1	3.4
Low-brain	63.8	5.74	51.9	81.4	4.36	75.9	3.61

New refers to a newborn NMR, 4 and 20 indicate the age of animals, and low indicates that samples were taken from an animal subjected to 8% O_2.

6.2 Gene expression levels

Gene expression levels were calculated as RPKM[32]. Transcriptome reads were mapped by Tophat, and the mapped reads were analyzed with in-house Perl scripts. To minimize the influence of difference in RNA output between the samples, the total read numbers were normalized by multiplying a normalization factor[33].

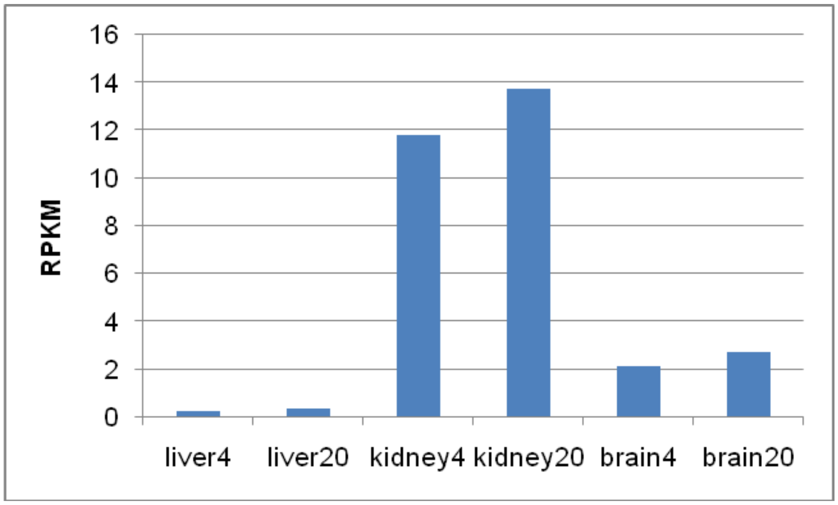
6.3 Differentially expressed genes and enrichment analysis

Differentially expressed genes were detected using the method of Chen et al.[34], which is based on the

Poisson distribution[35] and normalization for differences in the RNA output size and sequencing depth between samples, as well as accounting for different gene length. Genes with RPKM>5 in at least one experiment, at least 2 fold difference (in RPKM) in two experiments, and having fdr < 0.05 were defined as differentially expressed genes. Enrichment analysis was done using EnrichPipeline[32].

7 Unique traits

7.1 Aging



Supplementary Fig. 12. Expression of *TERT* in liver, kidney and brain of 4-year-old and 20-year-old NMRs.

7.2 Thermogenesis

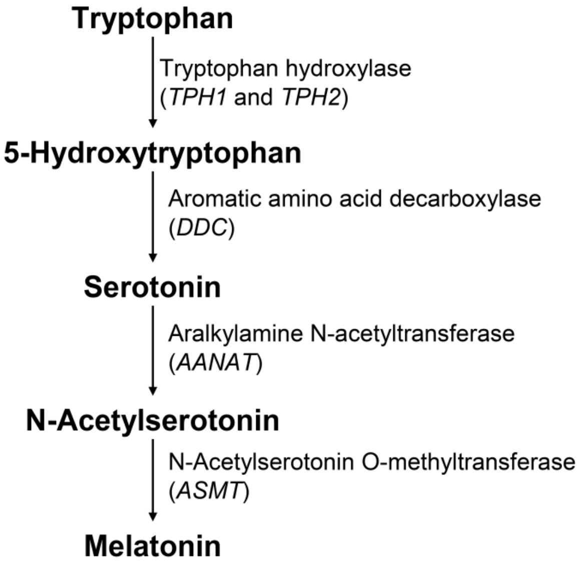
See Fig. 3 in the main text.

7.3 Melatonin pathways

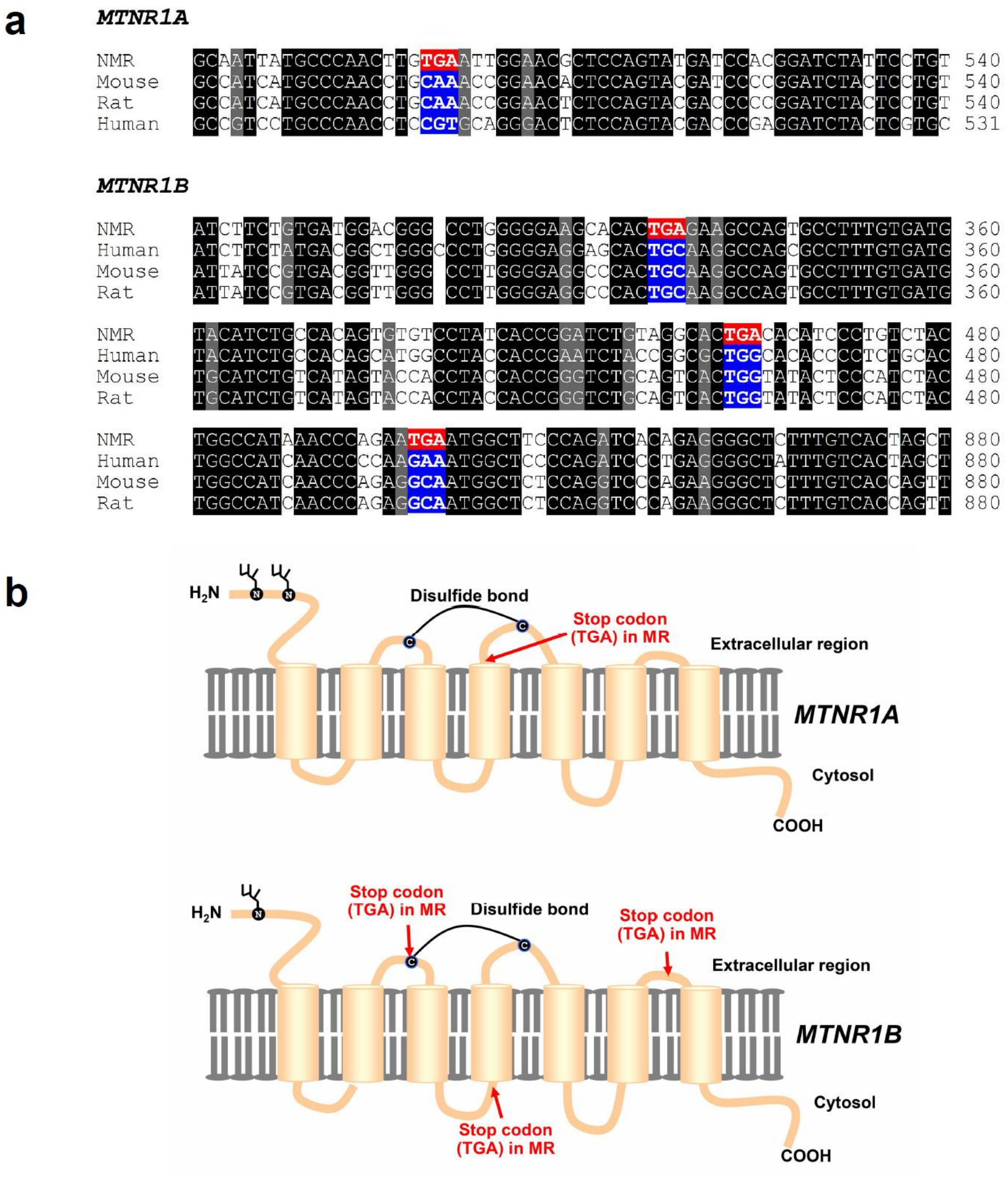
Supplementary Table 24. Expression of genes required for melatonin synthesis.

Genes	**Liver (age groups)**			**Kidney (age groups)**			**Brain (age groups)**		
	0	**4**	**20**	**0**	**4**	**20**	**0**	**4**	**20**
TPH1	0.05	0	0.24	1.13	0.31	0.33	1.28	1.36	1.59
TPH2	0	0	0.05	0	0.35	0	3.75	1.9	0.63
DDC	3.97	20.88	13.67	0.89	18.29	20.25	2.7	0.99	0.96
AANAT	1.76	0.9	0.39	0.98	1.31	2.33	0.27	0	0
ASMT	0	0	0	0	0	0	0	0	0

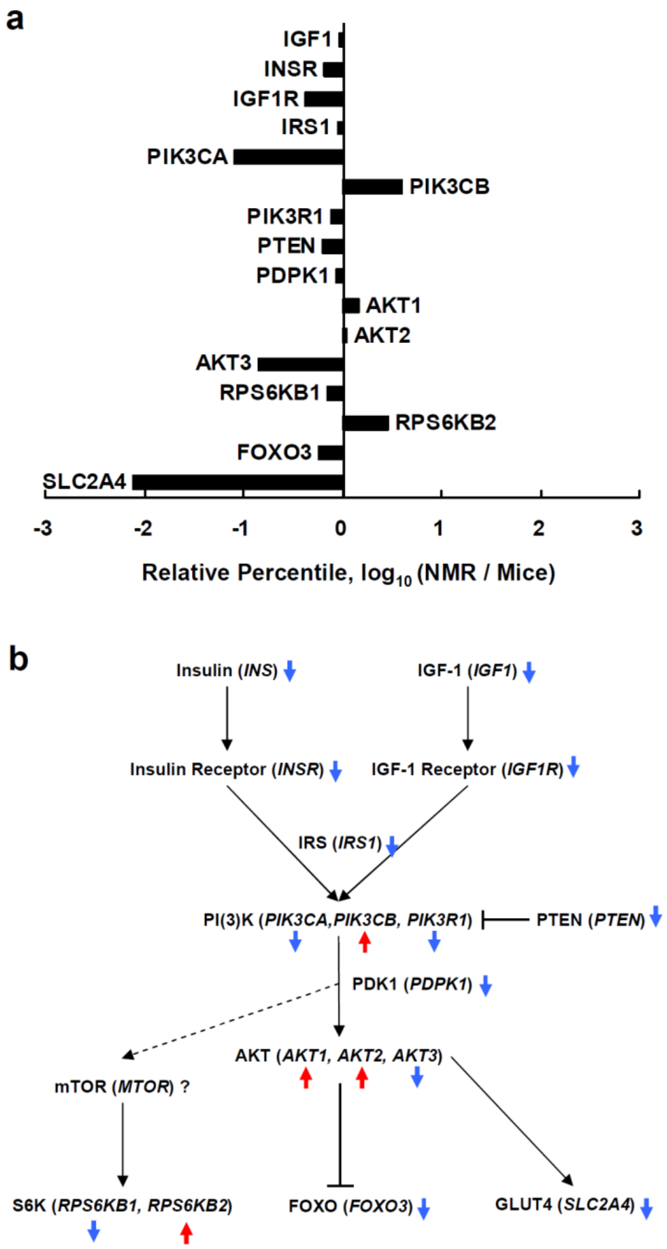
Gene expression is expressed as RPKM (Reads Per Kilo base-pair per Million mapped reads) for liver, kidney and brain for newborn (0), 4-year-old (4) and 20-year-old (20) NMRs.



Supplementary Fig. 13. Biosynthesis of melatonin. Compounds, genes and enzymes required for melatonin biosynthesis are shown.

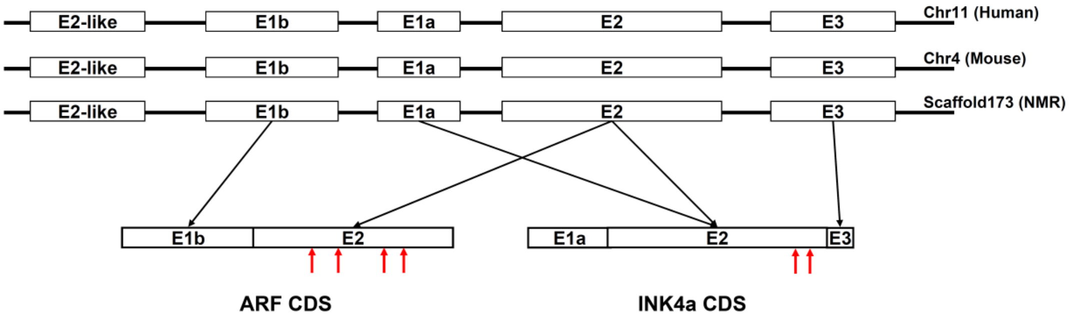

a
MTNR1A
NMR GCAATTATGCCCAACTTGTGAATTGGAACGCTCCAGTATGATCCACGGATCTATTCCTGT 540
Mouse GCCATCATGCCCAACCTGCAAACCGGAACACTCCAGTACGATCCCCGGATCTACTCCTGT 540
Rat GCCATCATGCCCAACCTGCAAACCGGAACTCTCCAGTACGACCCCCGGATCTACTCCTGT 540
Human GCCGTCCTGCCCAACCTCCGTGCAGGGACTCTCCAGTACGACCCGAGGATCTACTCGTGC 531
MTNR1B
NMR ATCTTCTGTGATGGACGGG CCTGGGGAAGCACACTGAGAAGCCAGTGCCTTTGTGATG 360
Human ATCTTCTATGACGGCTGGGCCCTGGGGGAGGAGCACTGCAAGGCCAGCGCCTTTGTGATG 360
Mouse ATTATCCGTGACGGTTGGG CCTTGGGGAGGCCCACTGCAAGGCCAGTGCCTTTGTGATG 360
Rat ATTATCCGTGACGGTTGGG CCTTGGGGAGGCCCACTGCAAGGCCAGTGCCTTTGTGATG 360
NMR TACATCTGCCACAGTGTGTCCTATCACCGGATCTGTAGGCACTGACACATCCCTGTCTAC 480
Human TACATCTGCCACAGCATGGCCTACCACCGAATCTACCGGCGCTGGCACACCCCTCTGCAC 480
Mouse TGCATCTGTCATAGTACCACCTACCACCGGGTCTGCAGTCACTGGTATACTCCCATCTAC 480
Rat TGCATCTGTCATAGTACCACCTACCACCGGGTCTGCAGTCACTGGTATACTCCCATCTAC 480
NMR TGGCCATAAACCCAGAATGAATGGCTTCCCAGATCACAGAGGGGCTCTTTGTCACTAGCT 880
Human TGGCCATCAACCCCCAAGAAATGGCTCCCCAGATCCCTGAGGGGCTATTTGTCACTAGCT 880
Mouse TGGCCATCAACCCAGAGGCAATGGCTCTCCAGGTCCCAGAAGGGCTCTTTGTCACCAGTT 880
Rat TGGCCATCAACCCAGAGGCAATGGCTCTCCAGGTCCCAGAAGGGCTCTTTGTCACCAGTT 880
b
H2N
N
N
Disulfide bond
C
C
Stop codon (TGA) in MR
Extracellular region
MTNR1A
Cytosol
COOH
H2N
N
Stop codon (TGA) in MR
Disulfide bond
C
C
Stop codon (TGA) in MR
Extracellular region
MTNR1B
Cytosol
Stop codon (TGA) in MR
COOH

Supplementary Fig. 14. Inactivation of melatonin receptors. (A) Premature stop codons (shown in red) in the NMR melatonin receptor sequences, *MTNR1A* and *MTNR1B*. (B) Topology of human melatonin receptors and the locations of stop codons in the NMR protein.

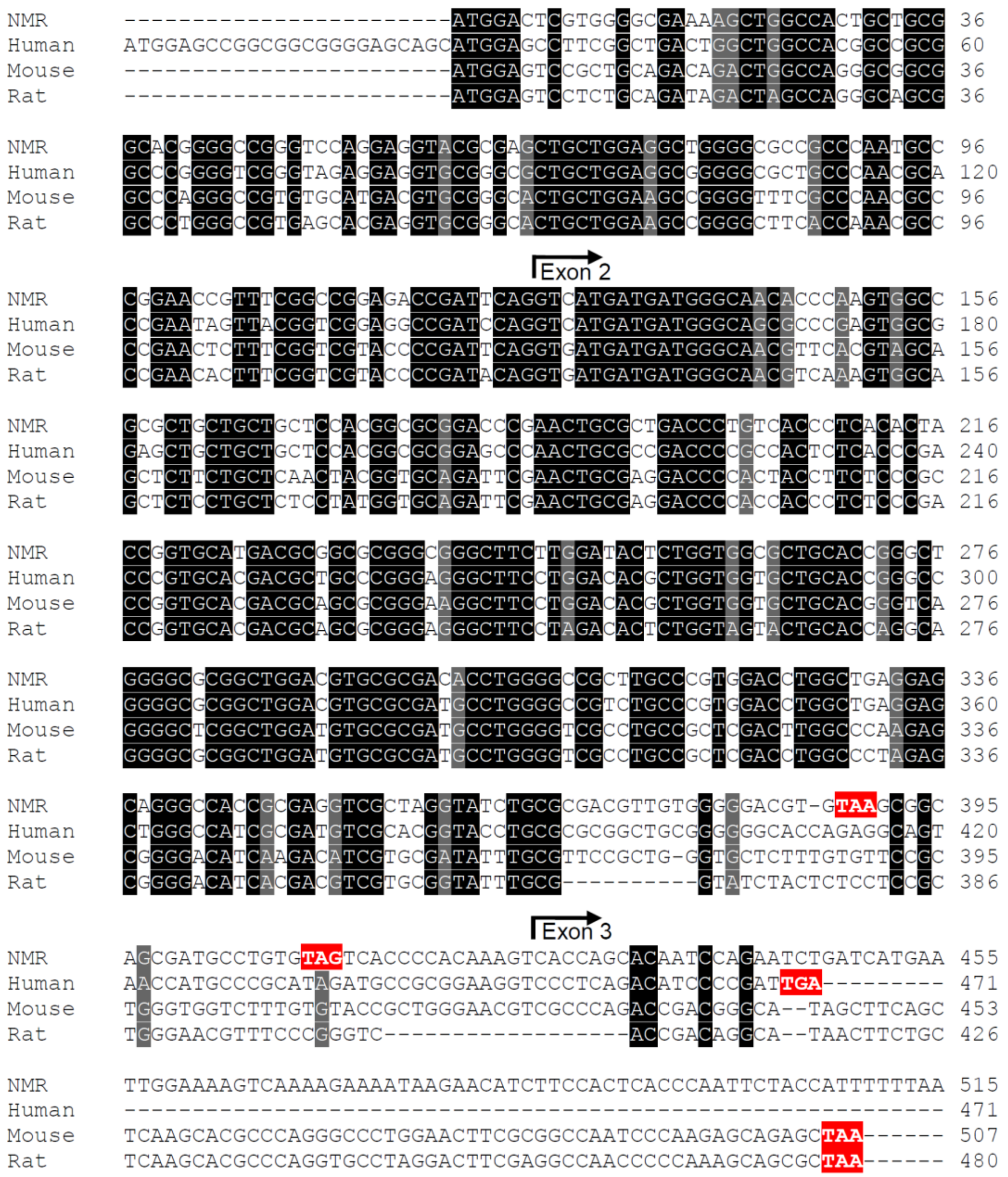


Supplementary Fig. 15. Down-regulation of insulin/IGF-1 signaling in the liver. (A) Decreased gene expression is expressed as relative percentile values, $\log_{10}$ (NMR/mice), based on the abundance in the NMR transcriptome data, average RPKM values from three individual NMR transcriptomes (4-20 year old) and mouse (10-12 week-old) microarray data from Gene Atlas[36]. (B) Down/up-regulation of genes in the insulin/IGF-1 signal pathway is shown with blue and red arrows, respectively, according to panel A.

7.4 Cancer

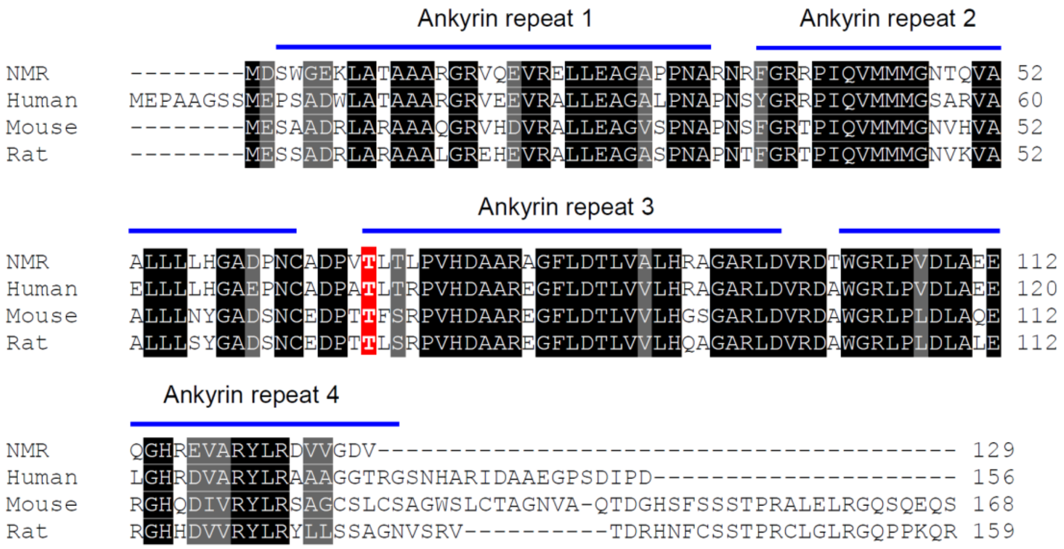


Supplementary Fig. 16. The *CDKN2A* locus within the genome. E1b, E1a, E2, and E3 are exons. The E2-like region is a sequence with homology to E2. Stop codons detected in NMR genes are indicated with red arrows.

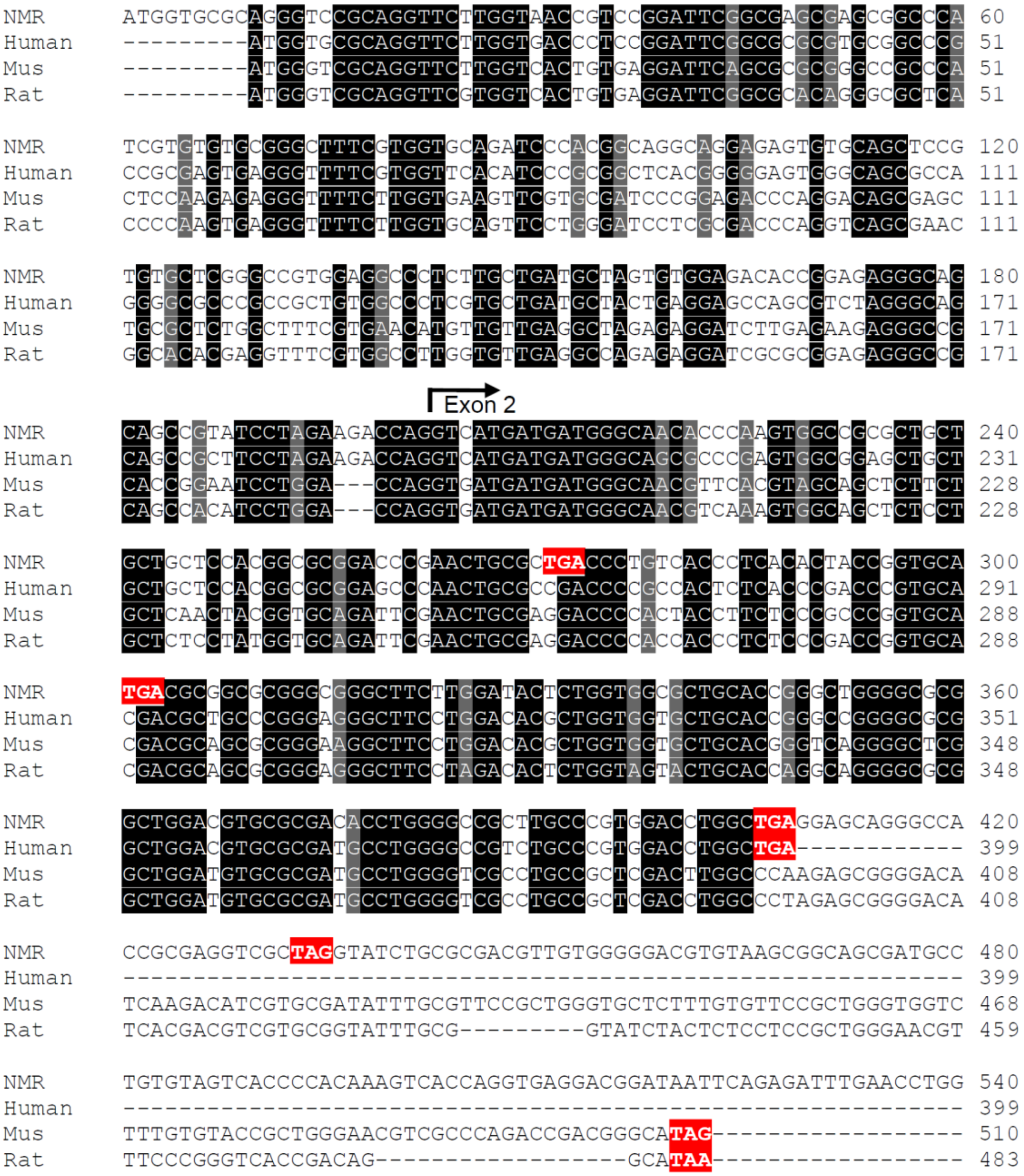


```
NMR      ------------------------ATGGACTCGTGGGGCGAAAAGCTGGCCACTGCTGCG 36
Human    ATGGAGCCGGCGGCGGGGAGCAGCATGGAGCCTTCGGCTGACTGGCTGGCCACGGCCGCG 60
Mouse    ------------------------ATGGAGTCCGCTGCAGACAGACTGGCCAGGGCGGCG 36
Rat      ------------------------ATGGAGTCCTCTGCAGATAGACTAGCCAGGGCAGCG 36

NMR      GCACGGGGCCGGGTCCAGGAGGTACGCGAGCTGCTGGAGGCTGGGGCGCCGCCCAATGCC 96
Human    GCCCGGGGTCGGGTAGAGGAGGTGCGGGCGCTGCTGGAGGCGGGGGCGCTGCCCAACGCA 120
Mouse    GCCCAGGGCCGTGTGCATGACGTGCGGGCACTGCTGGAAGCCGGGGTTTCGCCCAACGCC 96
Rat      GCCCTGGGCCGTGAGCACGAGGTGCGGGCACTGCTGGAAGCCGGGGCTTCACCAAACGCC 96

NMR      CGGAACCGTTTCGGCCGGAGACCGATTCAGGTCATGATGATGGGCAACACCCAAGTGGCC 156
Human    CCGAATAGTTACGGTCGGAGGCCGATCCAGGTCATGATGATGGGCAGCGCCCGAGTGGCG 180
Mouse    CCGAACTCTTTCGGTCGTACCCCGATTCAGGTGATGATGATGGGCAACGTTCACGTAGCA 156
Rat      CCGAACACTTTCGGTCGTACCCCGATACAGGTGATGATGATGGGCAACGTCAAAGTGGCA 156

NMR      GCGCTGCTGCTGCTCCACGGCGCGGACCCGAACTGCGCTGACCCTGTCACCCTCACACTA 216
Human    GAGCTGCTGCTGCTCCACGGCGCGGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGA 240
Mouse    GCTCTTCTGCTCAACTACGGTGCAGATTCGAACTGCGAGGACCCCACTACCTTCTCCCGC 216
Rat      GCTCTCCTGCTCTCCTATGGTGCAGATTCGAACTGCGAGGACCCCACCACCCTCTCCCGA 216

NMR      CCGGTGCATGACGCGGCGCGGGCGGGCTTCTTGGATACTCTGGTGGCGCTGCACCGGGCT 276
Human    CCCGTGCACGACGCTGCCCGGGAGGGCTTCCTGGACACGCTGGTGGTGCTGCACCGGGCC 300
Mouse    CCGGTGCACGACGCAGCGCGGGAAGGCTTCCTGGACACGCTGGTGGTGCTGCACGGGTCA 276
Rat      CCGGTGCACGACGCAGCGCGGGAGGGCTTCCTAGACACTCTGGTAGTACTGCACCAGGCA 276

NMR      GGGGCGCGGCTGGACGTGCGCGACACCTGGGGCCGCTTGCCCGTGGACCTGGCTGAGGAG 336
Human    GGGGCGCGGCTGGACGTGCGCGATGCCTGGGGCCGTCTGCCCGTGGACCTGGCTGAGGAG 360
Mouse    GGGGCTCGGCTGGATGTGCGCGATGCCTGGGGTCGCCTGCCGCTCGACTTGGCCCAAGAG 336
Rat      GGGGCGCGGCTGGATGTGCGCGATGCCTGGGGTCGCCTGCCGCTCGACCTGGCCCTAGAG 336

NMR      CAGGGCCACCGCGAGGTCGCTAGGTATCTGCGCGACGTTGTGGGGGACGT-GTAAGCGGC 395
Human    CTGGGCCATCGCGATGTCGCACGGTACCTGCGCGCGGCTGCGGGGGGCACCAGAGGCAGT 420
Mouse    CGGGGACATCAAGACATCGTGCGATATTTGCGTTCCGCTG-GGTGCTCTTTGTGTTCCGC 395
Rat      CGGGGACATCACGACGTCGTGCGGTATTTGCG----------GTATCTACTCTCCTCCGC 386

NMR      AGCGATGCCTGTGTAGTCACCCCACAAAGTCACCAGCACAATCCAGAATCTGATCATGAA 455
Human    AACCATGCCCGCATAGATGCCGCGGAAGGTCCCTCAGACATCCCCGATTGA--------- 471
Mouse    TGGGTGGTCTTTGTGTACCGCTGGGAACGTCGCCCAGACCGACGGGCA--TAGCTTCAGC 453
Rat      TGGGAACGTTTCCCGGGTC------------------ACCGACAGGCA--TAACTTCTGC 426

NMR      TTGGAAAAGTCAAAAGAAAATAAGAACATCTTCCACTCACCCAATTCTACCATTTTTTAA 515
Human    ------------------------------------------------------------ 471
Mouse    TCAAGCACGCCCAGGGCCCTGGAACTTCGCGGCCAATCCCAAGAGCAGAGCTAA------ 507
Rat      TCAAGCACGCCCAGGTGCCTAGGACTTCGAGGCCAACCCCCAAAGCAGCGCTAA------ 480
```

Supplementary Fig. 17. Alignment of mammalian Ink4a coding regions. Stop codons are shown in red.

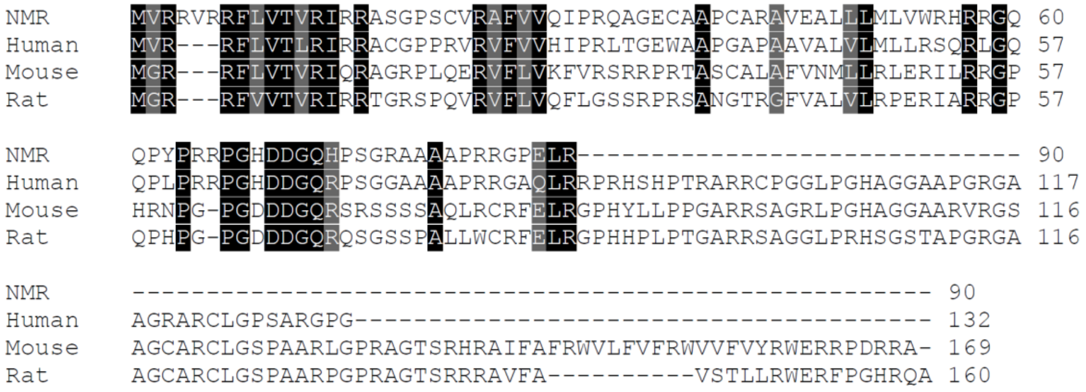

Ankyrin repeat 1
Ankyrin repeat 2
NMR --------MDSWGEKLATAAARGRVQEVRELLEAGAPPNARNRFGRRPIQVMMMGNTQVA 52
Human MEPAAGSSMEPSADWLATAAARGRVEEVRALLEAGALPNAPNSYGRRPIQVMMMGSARVA 60
Mouse --------MESAADRLARAAAQGRVHDVRALLEAGVSPNAPNSFGRTPIQVMMMGNVHVA 52
Rat --------MESSADRLARAAALGREHEVRALLEAGASPNAPNTFGRTPIQVMMMGNVKVA 52
Ankyrin repeat 3
NMR ALLLLHGADPNCADPVTLTLPVHDAARAGFLDTLVALHRAGARLDVRDTWGRLPVDLAEE 112
Human ELLLLHGAEPNCADPATLTRPVHDAAREGFLDTLVVLHRAGARLDVRDAWGRLPVDLAEE 120
Mouse ALLLNYGADSNCEDPTTFSRPVHDAAREGFLDTLVVLHGSGARLDVRDAWGRLPLDLAQE 112
Rat ALLLSYGADSNCEDPTTLSRPVHDAAREGFLDTLVVLHQAGARLDVRDAWGRLPLDLALE 112
Ankyrin repeat 4
NMR QGHREVARYLRDVVGDV-- 129
Human LGHRDVARYLRAAAGGTRGSNHARIDAAEGPSDIPD--------------------- 156
Mouse RGHQDIVRYLRSAGCSLCSAGWSLCTAGNVA-QTDGHSFSSSTPRALELRGQSQEQS 168
Rat RGHHDVVRYLRYLLSSAGNVSRV----------TDRHNFCSSTPRCLGLRGQPPKQR 159

Supplementary Fig. 18. Alignment of mammalian Ink4a protein sequences. Location of four conserved ankyrin repeats is shown by blue lines above the sequence. The conserved threonine residue important for CDK6 binding is shown in red.



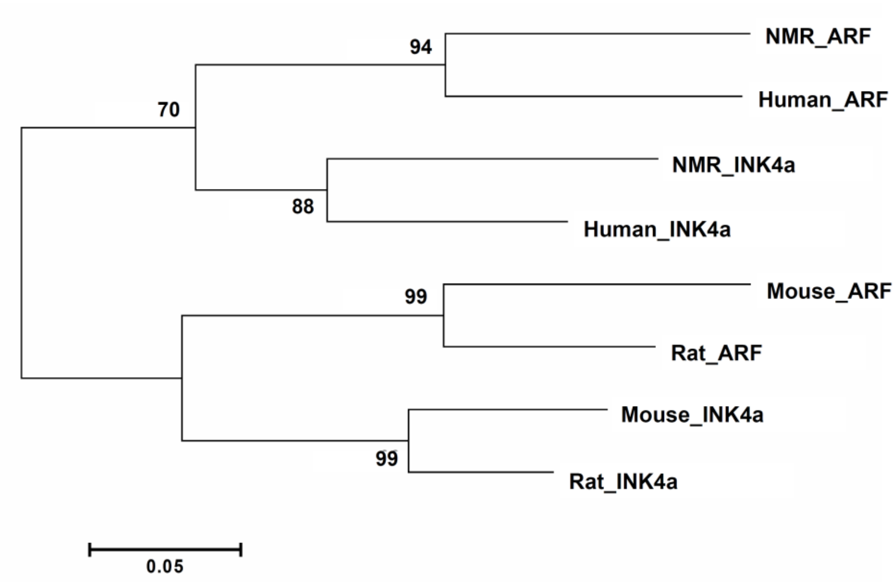
```
NMR      ATGGTGCGCAGGTCCGCAGGTTCTTGGTAACCGTCCGGATTCGGCGAGCGAGCGGCCCA 60
Human    ---------ATGGTGCGCAGGTTCTTGGTGACCCTCCGGATTCGGCGCGCGTGCGGCCCG 51
Mus      ---------ATGGGTCGCAGGTTCTTGGTCACTGTGAGGATTCAGCGCGCGGGCCGCCCA 51
Rat      ---------ATGGGTCGCAGGTTCGTGGTCACTGTGAGGATTCGGCGCACAGGGCGCTCA 51

NMR      TCGTGTGTGCGGGCTTTCGTGGTGCAGATCCCACGGCAGGCAGGAGAGTGTGCAGCTCCG 120
Human    CCGCGAGTGAGGGTTTTCGTGGTTCACATCCCGCGGCTCACGGGGAGTGGGCAGCGCCA 111
Mus      CTCCAAGAGAGGGTTTTCTTGGTGAAGTTCGTGCGATCCCGGAGACCCAGGACAGCGAGC 111
Rat      CCCCAAGTGAGGGTTTTCTTGGTGCAGTTCCTGGGATCCTCGCGACCCAGGTCAGCGAAC 111

NMR      TGTGCTCGGGCCGTGGAGGCCCTCTTGCTGATGCTAGTGTGGAGACACCGGAGAGGGCAG 180
Human    GGGGCGCCCGCCGCTGTGGCCCTCGTGCTGATGCTACTGAGGAGCCAGCGTCTAGGGCAG 171
Mus      TGCGCTCTGGCTTTCGTGAACATGTTGTTGAGGCTAGAGAGGATCTTGAGAAGAGGGCCG 171
Rat      GGCACACGAGGTTTCGTGGCCTTGGTGTTGAGGCCAGAGAGGATCGCGCGGAGAGGGCCG 171

NMR      CAGCCGTATCCTAGAAGACCAGGTCATGATGATGGGCAACACCCAAGTGGCCGCGCTGCT 240
Human    CAGCCGCTTCCTAGAAGACCAGGTCATGATGATGGGCAGCGCCCGAGTGGCGGAGCTGCT 231
Mus      CACCGGAATCCTGGA---CCAGGTGATGATGATGGGCAACGTTCACGTAGCAGCTCTTCT 228
Rat      CAGCCACATCCTGGA---CCAGGTGATGATGATGGGCAACGTCAAAGTGGCAGCTCTCCT 228

NMR      GCTGCTCCACGGCGCGGACCCGAACTGCGCTGACCCTGTCACCCTCACACTACCGGTGCA 300
Human    GCTGCTCCACGGCGCGGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGACCCGTGCA 291
Mus      GCTCAACTACGGTGCAGATTCGAACTGCGAGGACCCCACTACCTTCTCCCGCCCGGTGCA 288
Rat      GCTCTCCTATGGTGCAGATTCGAACTGCGAGGACCCCACCACCCTCTCCCGACCGGTGCA 288

NMR      TGACGCGGCGCGGGCGGGCTTCTTGGATACTCTGGTGGCGCTGCACCGGGCTGGGGCGCG 360
Human    CGACGCTGCCCGGGAGGGCTTCCTGGACACGCTGGTGGTGCTGCACCGGGCCGGGGCGCG 351
Mus      CGACGCAGCGCGGGAAGGCTTCCTGGACACGCTGGTGGTGCTGCACGGGTCAGGGGCTCG 348
Rat      CGACGCAGCGCGGGAGGGCTTCCTAGACACTCTGGTAGTACTGCACCAGGCAGGGGCGCG 348

NMR      GCTGGACGTGCGCGACACCTGGGGCCGCTTGCCCGTGGACCTGGCTGAGGAGCAGGGCCA 420
Human    GCTGGACGTGCGCGATGCCTGGGGCCGTCTGCCCGTGGACCTGGCTGA------------ 399
Mus      GCTGGATGTGCGCGATGCCTGGGGTCGCCTGCCGCTCGACTTGGCCCAAGAGCGGGGACA 408
Rat      GCTGGATGTGCGCGATGCCTGGGGTCGCCTGCCGCTCGACCTGGCCCTAGAGCGGGGACA 408

NMR      CCGCGAGGTCGCTAGGTATCTGCGCGACGTTGTGGGGGACGTGTAAGCGGCAGCGATGCC 480
Human    ------------------------------------------------------------ 399
Mus      TCAAGACATCGTGCGATATTTGCGTTCCGCTGGGTGCTCTTTGTGTTCCGCTGGGTGGTC 468
Rat      TCACGACGTCGTGCGGTATTTGCG---------GTATCTACTCTCCTCCGCTGGGAACGT 459

NMR      TGTGTAGTCACCCCACAAAGTCACCAGGTGAGGACGGATAATTCAGAGATTTGAACCTGG 540
Human    ------------------------------------------------------------ 399
Mus      TTTGTGTACCGCTGGGAACGTCGCCCAGACCGACGGGCATAG------------------ 510
Rat      TTCCCGGGTCACCGACAG------------------GCATAA------------------ 483
```

Supplementary Fig. 19. Alignment of mammalian Arf coding regions. Stop codons are shown in red.



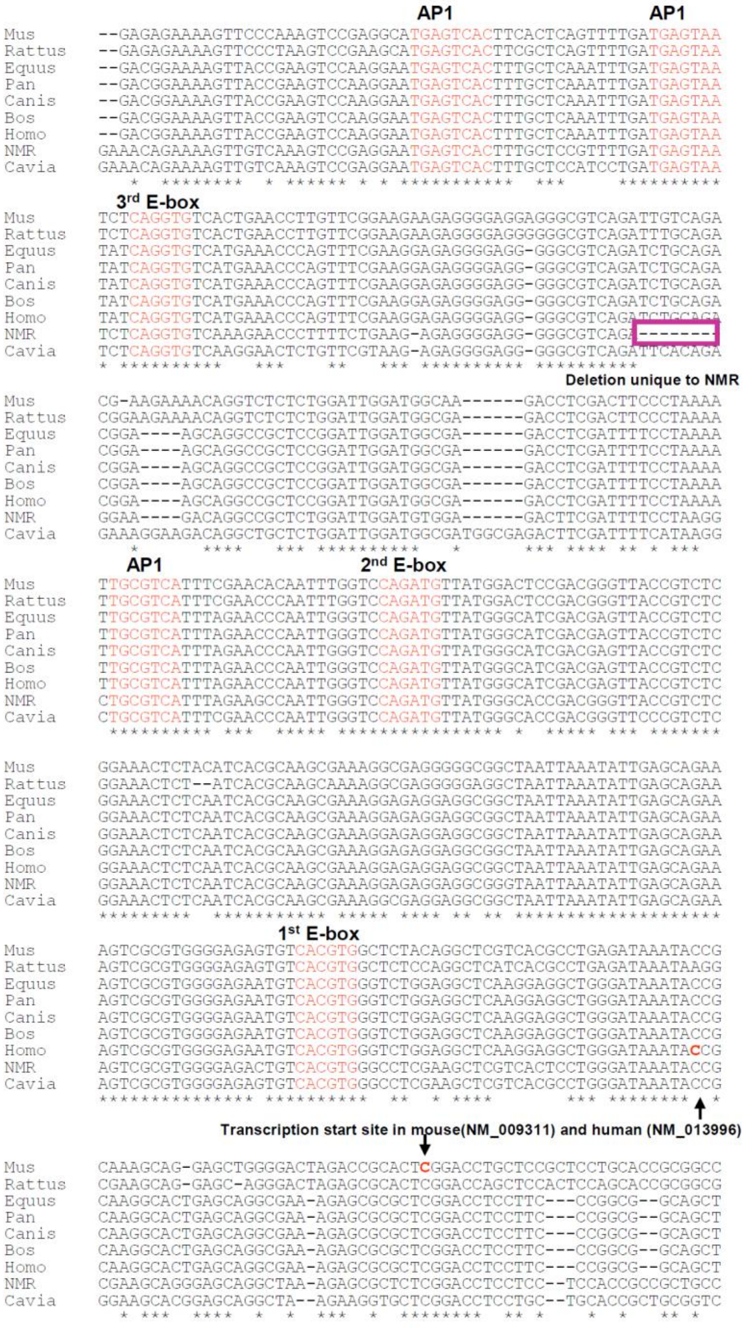
```
NMR      MVRRVRRFLVTVRIRRASGPSCVRAFVVQIPRQAGECAAPCARAVEALLLMLVWRHRRGQ 60
Human    MVR---RFLVTLRIRRACGPPRVRVFVVHIPRLTGEWAAPGAPAAVALVLMLLRSQRLGQ 57
Mouse    MGR---RFLVTVRIQRAGRPLQERVFLVKFVRSRRPRTASCALAFVNMLLRLERILRRGP 57
Rat      MGR---RFVVTVRIRRTGRSPQVRVFLVQFLGSSRPRSANGTRGFVALVLRPERIARRGP 57

NMR      QPYPRRPGHDDGQHPSGRAAAAPRRGPELR------------------------------ 90
Human    QPLPRRPGHDDGQRPSGGAAAAPRRGAQLRRPRHSHPTRARRCPGGLPGHAGGAAPGRGA 117
Mouse    HRNPG-PGDDDGQRSRSSSSAQLRCRFELRGPHYLLPPGARRSAGRLPGHAGGAARVRGS 116
Rat      QPHPG-PGDDDGQRQSGSSPALLWCRFELRGPHHPLPTGARRSAGGLPRHSGSTAPGRGA 116

NMR      ------------------------------------------------------ 90
Human    AGRARCLGPSARGPG--------------------------------------- 132
Mouse    AGCARCLGSPAARLGPRAGTSRHRAIFAFRWVLFVFRWVVFVYRWERRPDRRA- 169
Rat      AGCARCLGSPAARPGPRAGTSRRRAVFA----------VSTLLRWERFPGHRQA 160
```

Supplementary Fig. 20. Alignment of mammalian Arf protein sequences.



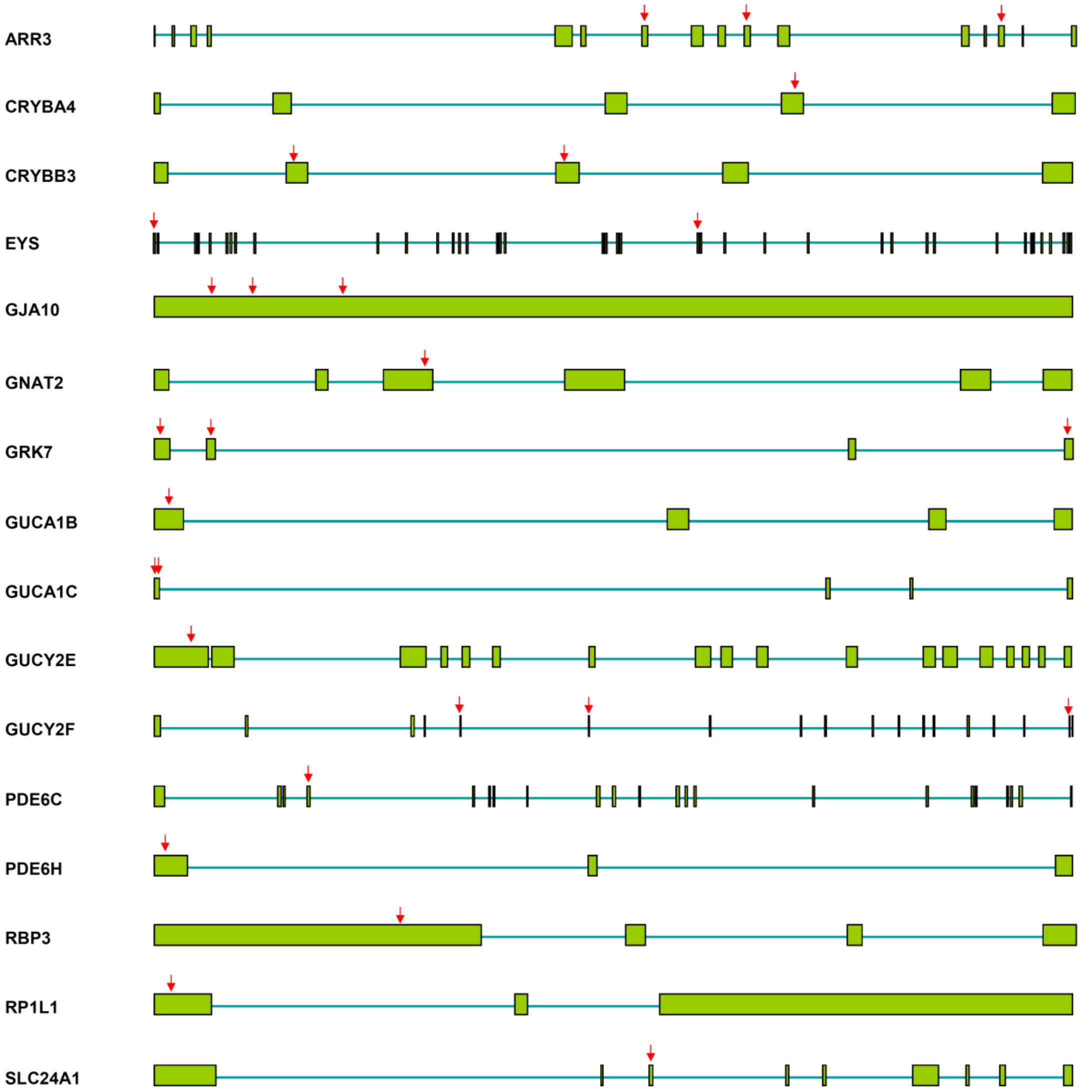
Supplementary Fig. 21. Phylogenetic tree of Ink4a and Arf coding regions. Scale bar shows sequence divergence (0.05 = 5 %).

7.5 Pain sensitivity



Supplementary Fig 22. NMR-specific deletion within the *TAC1* promoter. Transcription start sites of human and mouse genes are indicated with arrows. NMR-specific deletion within the *TAC1* promoter is indicated with a box. AP1 and E-box are transcription factor binding sites known to regulate *TAC1* expression.

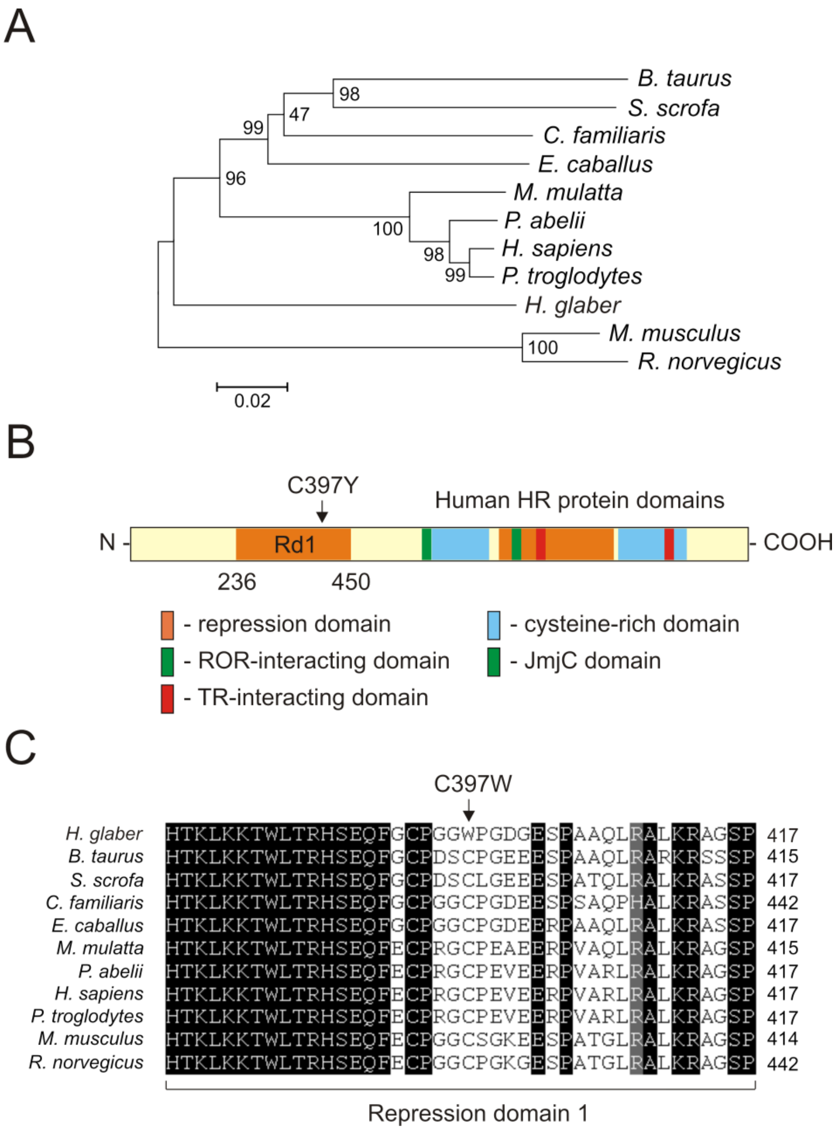
7.6 Visual function



Supplementary Fig. 23. Inactivation events identified within visual perception pseudogenes. Gene structures are based on the orthologous mouse or human genes. Green squares indicate exons and blue lines introns. Red arrows show inactivation events, such as insertion or deletion that change the frame, or point mutations resulting in premature termination.

7.7 Hairless phenotype

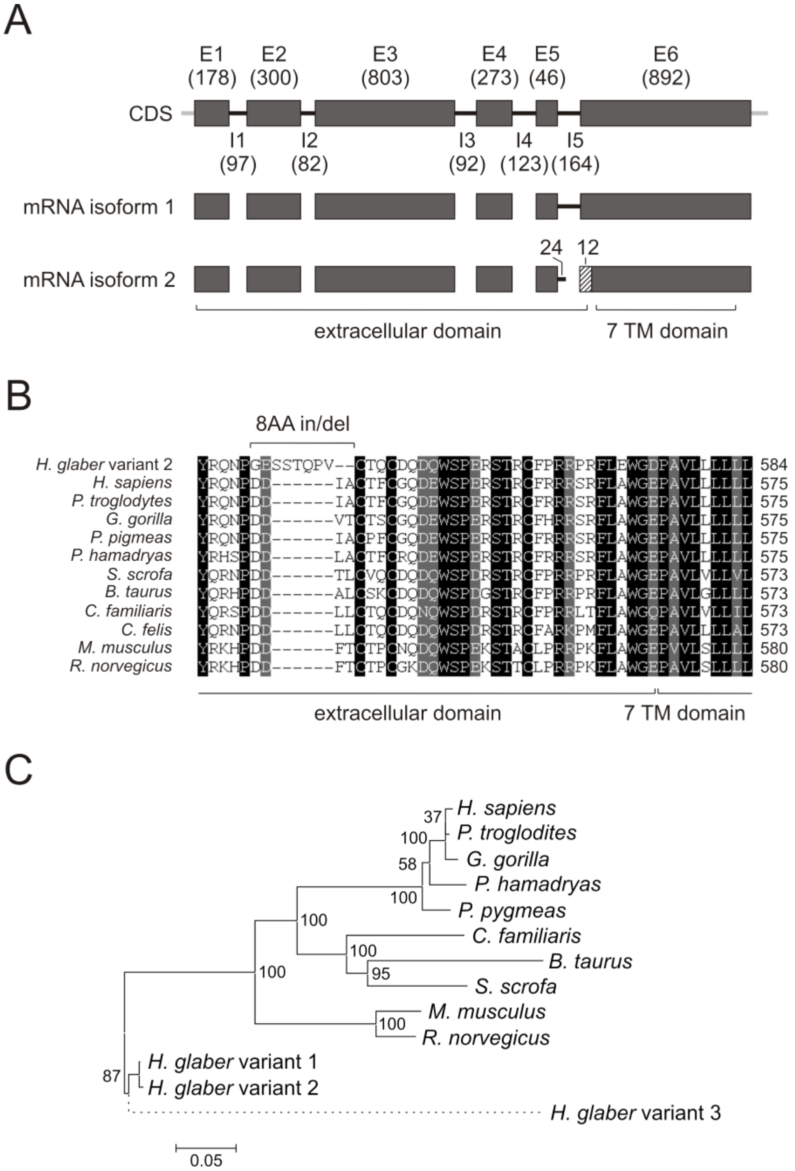
NMR is the only mole rat species that naturally has no fur; however, these animals do have about 100 sensory hairs on their bodies that act like whiskers to help feel the surroundings. In mammals, much of the understanding of the molecular pathways of hair growth has come from the studies on the function of the nuclear receptor co-repressor, Hairless (Hr), whose mutations cause hair loss in mice, rat and men. Analyses of NMR Hr revealed substantial divergence of this protein from known mammalian orthologs and the presence of mutations specifically associated with the hairless phenotype (Supplementary Fig. 24).


A
B. taurus
S. scrofa
C. familiaris
E. caballus
M. mulatta
P. abelii
H. sapiens
P. troglodytes
H. glaber
M. musculus
R. norvegicus
98
47
99
96
100
98
99
100
0.02
B
C397Y
Human HR protein domains
N -
Rd1
- COOH
236
450
- repression domain
- ROR-interacting domain
- TR-interacting domain
- cysteine-rich domain
- JmjC domain
C
C397W
H. glaber HTKLKKTWLTRHSEQFGCPGGWPGDGESPAAQLRALKRAGSP 417
B. taurus HTKLKKTWLTRHSEQFGCPDSCPGEEESPAAQLRARKRSSSP 415
S. scrofa HTKLKKTWLTRHSEQFGCPDSCLGEEESPATQLRALKRASSP 417
C. familiaris HTKLKKTWLTRHSEQFGCPGGCPGDEESPSAQPHALKRASSP 442
E. caballus HTKLKKTWLTRHSEQFGCPGGCPGDEERPAAQLRALKRASSP 417
M. mulatta HTKLKKTWLTRHSEQFECPRGCPEAEERPVAQLRALKRAGSP 415
P. abelii HTKLKKTWLTRHSEQFECPRGCPEVEERPVARLRALKRAGSP 417
H. sapiens HTKLKKTWLTRHSEQFECPRGCPEVEERPVARLRALKRAGSP 417
P. troglodytes HTKLKKTWLTRHSEQFECPRGCPEVEERPVARLRALKRAGSP 417
M. musculus HTKLKKTWLTRHSEQFECPGGCSGKEESPATGLRALKRAGSP 414
R. norvegicus HTKLKKTWLTRHSEQFECPGGCPGKGESPATGLRALKRAGSP 442
Repression domain 1

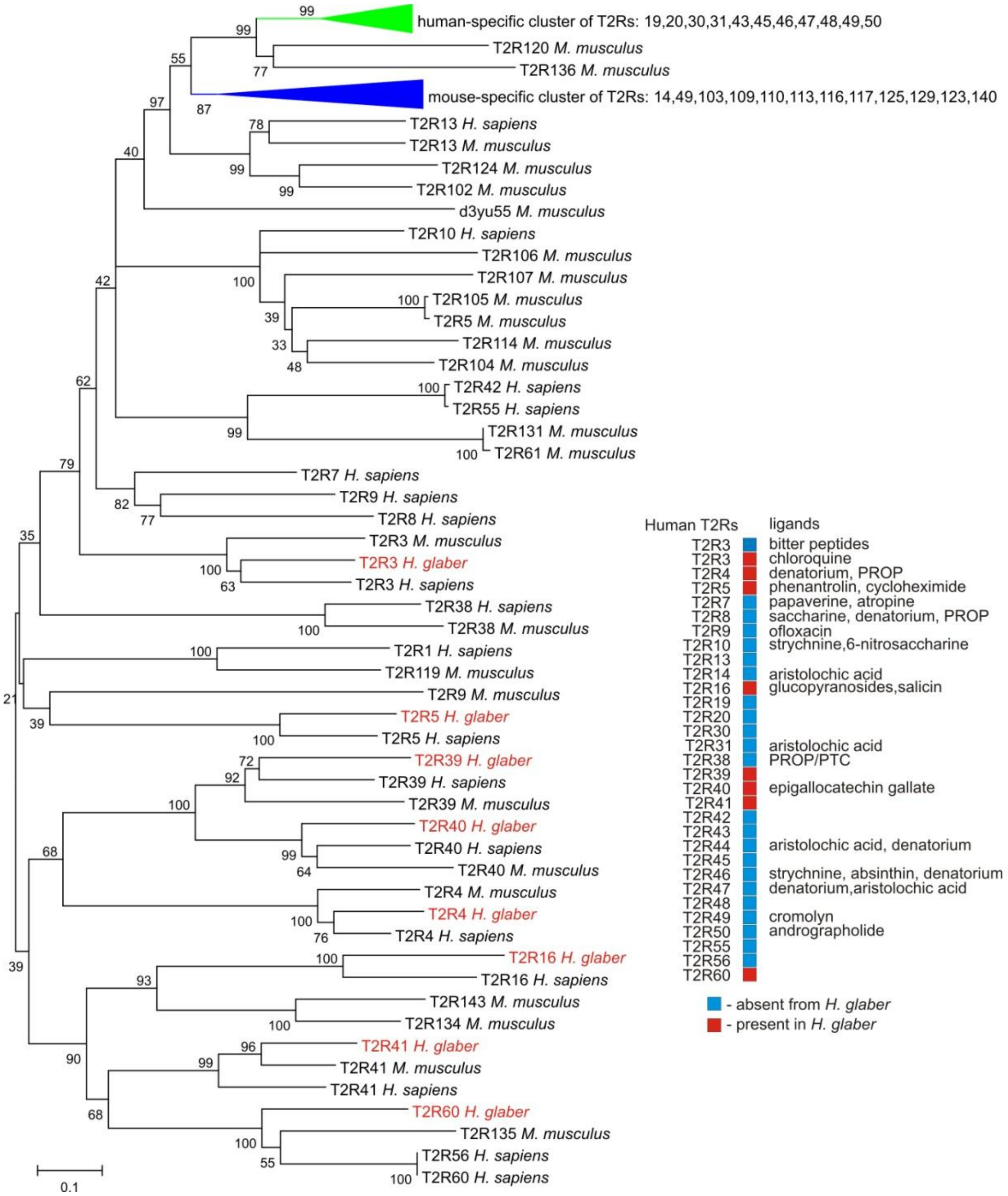
Supplementary Fig. 24. Hairless homolog (Hr) of *H. glaber*. (A) The Neighbor-Joining tree demonstrating relationships between the predicted NMR Hr homolog protein and ten Hr proteins from other mammals: *R. norvegicus* (NP_077340.2), *M. musculus* (NP_068677.2), *P. troglodites* (XP_001153297.1), *H. sapiens* (NP_005135.2), *M. mulata* (NP_001028015.1), *E. caballus* (XP_001490941.2), *S. scrofa* (NP_001077399.1), *B. taurus* (NP_001096005.1), *P. abelii* (XP_002818913), *C. familiaris* (XP_543256). The percentage of replicate trees, in which the associated taxa clustered together in the bootstrap test (1000 replicates), is shown next to the branches. (B) Schematic representation of human HR protein functional domains. Repression domains (RD1, 236-450; RD2, 750-864; RD3, 864-981); TR-interacting domains (TR-ID1, 816-830; TR-ID2, 1026-1038); ROR-interacting domains, ROR-ID1, 586-590; ROR-ID2, 778-782); cysteine-rich domain, 587-712; JmjC domain, 964-1175. Note that rat Hr is 1207 amino acids; mouse and human Hr initiate at an internal AUG (amino acid 27 in rat Hr) and are 1182 and 1189 amino acids, respectively. Rat hairless phenotype polymorphism: C397Y/C422Y. (C) Protein alignment of Hr proteins from NMR and ten other mammals. The position of C397Y mutation associated with rat hairless phenotype is indicated on the top. In the NMR sequence, this Cys is replaced with Trp.

7.8 Senses of taste

Complex sense of taste developed in the animal kingdom as a mechanism to survive in the environments featuring millions of different compounds. In NMR, we observed a substantial sequence variation (in respect to other mammals) in *T1R3*, a common component of heterodimeric sweet and umami taste receptor, resulted in alternative splicing and several mRNA isoforms (Supplementary Fig. 25). The isoform 1 encodes a protein lacking all transmembrane helices, suggesting a loss of protein function. Another mRNA isoform 2 leads to an uncommon exchange between the extracellular and transmembrane protein segments (Supplementary Fig. 25). Since even a single amino acid replacement in *T1R3* can alter protein function, the data suggest an altered function of sweet taste in NMR. In addition, we have detected only eight bitter taste receptors in the NMR genome, which is much lower than in human (32 receptors) and mouse (36 receptors) (Supplementary Fig. 26).


A
E1
(178)
E2
(300)
E3
(803)
E4
(273)
E5
(46)
E6
(892)
CDS
I1
(97)
I2
(82)
I3
(92)
I4
(123)
I5
(164)
mRNA isoform 1
24
12
mRNA isoform 2
extracellular domain
7 TM domain
B
8AA in/del
H. glaber variant 2 YRQNPGESSTQPV--CTQCDQDQWSPERSTRCFPRRPRFLEWGDPAVLLLLLL 584
H. sapiens YRQNPDD------IACTFCGQDEWSPERSTRCFRRRSRFLAWGEPAVLLLLLL 575
P. troglodytes YRQNPDD------IACTFCGQDEWSPERSTRCFRRRSRFLAWGEPAVLLLLLL 575
G. gorilla YRQNPDD------VTCTSCGQDEWSPERSTRCFHRRSRFLAWGEPAVLLLLLL 575
P. pigmeas YRQNPDD------IACPFCGQDEWSPERSTRCFRRRSRFLAWGEPAVLLLLLL 575
P. hamadryas YRHSPDD------LACTFCRQDEWSPERSTRCFRRRSRFLAWGEPAVLLLLLL 575
S. scrofa YQRNPDD------TLCVQCDQDQWSPDRSTRCFPRRPRFLAWGEPAVLVLLVL 573
B. taurus YQRHPDD------ALCSKCDQDQWSPDGSTRCFPRRPRFLAWGEPAVLGLLLL 573
C. familiaris YQRSPDD------LLCTQCDQNQWSPDRSTRCFPRRLTFLAWGQPAVLVLLIL 573
C. felis YQRNPDD------LLCTQCDQDQWSPDRSTRCFARKPMFLAWGEPAVLLLLAL 573
M. musculus YRKHPDD------FTCTPCNQDQWSPEKSTACLPRRPKFLAWGEPVVLSLLLL 580
R. norvegicus YRKHPDD------FTCTPCGKDQWSPEKSTTCLPRRPKFLAWGEPAVLSLLLL 580
extracellular domain
7 TM domain
C
37
H. sapiens
100
P. troglodites
58
G. gorilla
P. hamadryas
100
P. pygmeas
100
C. familiaris
100
B. taurus
100
95
S. scrofa
M. musculus
100
R. norvegicus
H. glaber variant 1
87
H. glaber variant 2
H. glaber variant 3
0.05

Supplementary Fig. 25. The genetic structure of *H. glaber Tas1r3* and its phylogenetic relationship with other mammals. (A) Coding sequence (CDS) and two mRNA isoforms of NMR *Tas1r3*. The sizes of exons (E) and introns (I) are in bp. Isoform 1 contains an unspliced intron I5 between E5 and E6. mRNA isoform 2 is the product of an alternative splicing event, harbors additional 25 bp of I5, and lacks 12 bp of exon 6 leading to an uncommon protein insertion/deletion variant. The encoded extracellular and 7-transmembrane protein domains are indicated on the bottom. (B) Multiple sequence alignment of T1R3 proteins from 11 mammals and the *H. glaber* protein variant 2. (C) The optimal Neighbor-Joining phylogenetic tree demonstrating the relationship between the *H. glaber* T1R3 protein variants and known T1R3 proteins in other mammals. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset.



Supplementary Fig. 26. The Neighbor-Joining phylogenetic tree demonstrating the relationships between eight NMR T2R proteins (in red) and known T2R proteins of human and mouse. The cluster in green corresponds to eleven human-specific T2Rs, and the cluster in blue contains twelve mouse-specific T2Rs. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset.

7.9 Gene expression analyses of NMR subjected to low O_2

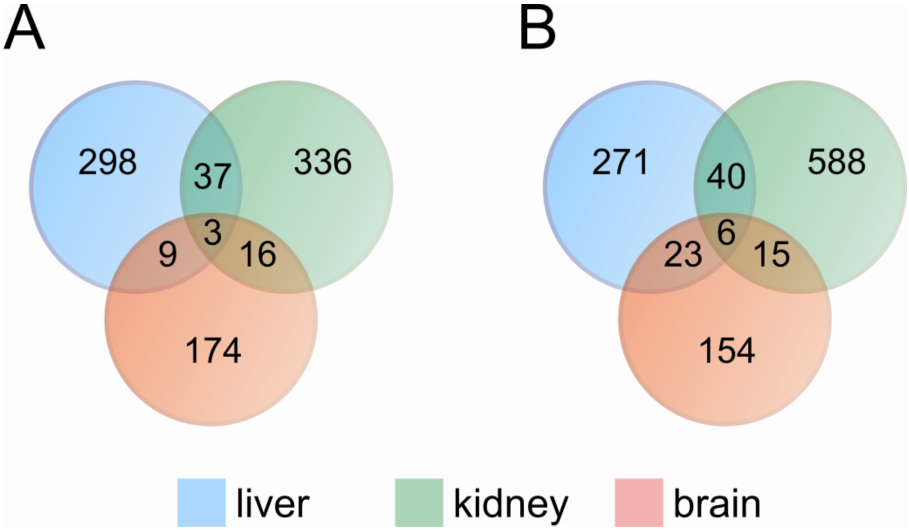
For this experiment, adult non-breeding NMRs (4-year old, female) were used. The experiment consisted of keeping the animals for a total period of 192 h, either in a chamber flushed with a mixture of oxygen and nitrogen (8:92) or in a regular housing chamber. After 192 h, the animals were sacrificed and whole brains, kidney and livers were rapidly removed, flash-frozen in liquid nitrogen, and stored at -80 °C until use. The whole transcriptome sequencing and calculation of gene expression levels (RPKM) for normoxic and hypoxic liver, kidney and brain were performed as described above. We used three technical replicates for each RNA sample. The replicated gene counts were highly correlated across the lanes (average Spearman correlation = 0.96). For each gene, we computed a goodness-of-fit statistic across 3 lanes to test the hypothesis: if there is no lane effect, then this statistic should be $\chi2$ distributed on $L - 1$ degrees of freedom. For lanes sequencing the same sample at the same concentration, only a small proportion of genes showed evidence of differences among lanes over those expected from sampling error. Those genes were removed from the dataset prior to the analysis. For the remaining genes, the average RPKMs were produced using three technical replicates. Although RNA-Seq is not affected by background from cross-hybridization, as microarrays are, it is not free of ambiguities caused by instrumental detection errors and amount of RNA used. Therefore, we applied robust quantile-based approach to normalize the RPKMs between each pair of normoxic/hypoxic tissues (assuming that at least 70% of gene expression remained unchanged between the normoxia and hypoxia samples). Low abundant genes with expression values <10 RPKM were removed from the dataset prior to analyses. To identify differentially expressed genes from the Illumina sequencing data, we compared the averaged RPKMs of normoxia/hypoxia liver, kidney and brain samples. At an FDR of 0.01%, we identified 661, 1003 and 382 genes as differentially expressed for the normoxia/hypoxia liver, kidney and brain samples, respectively (100% of these had an estimated absolute fold change >2). A quantile-quantile based approach was used to identify the point when the observed fold change starts deviating from the expected values under the null hypothesis of no changes in gene expression.

Supplementary Tables 25-31 and Supplementary Fig. 27 provide an overview of NMR genes which significantly (>2 fold) change their expression levels in brain, liver and kidney in response to the low oxygen atmosphere. Supplementary Fig. 28 summarizes putative functional consequences of these changes on NMR metabolism. In the liver, lower gene expression in low O_2 conditions was associated with energy metabolism, particularly homeostasis of triglycerides and lipids (Supplementary Table 25 and Supplementary Fig. 29). In addition, enrichment for GO terms was observed for sterol and cholesterol biosynthesis. Numerous upregulated genes in the liver were involved in immune signaling (including chemokine and IFN-γ pathways), supporting a link between immune function and hypoxia (Supplementary Table 28). A significant fraction of upregulated genes was related to the iron transport, apoptosis, and defense against hydroperoxides.

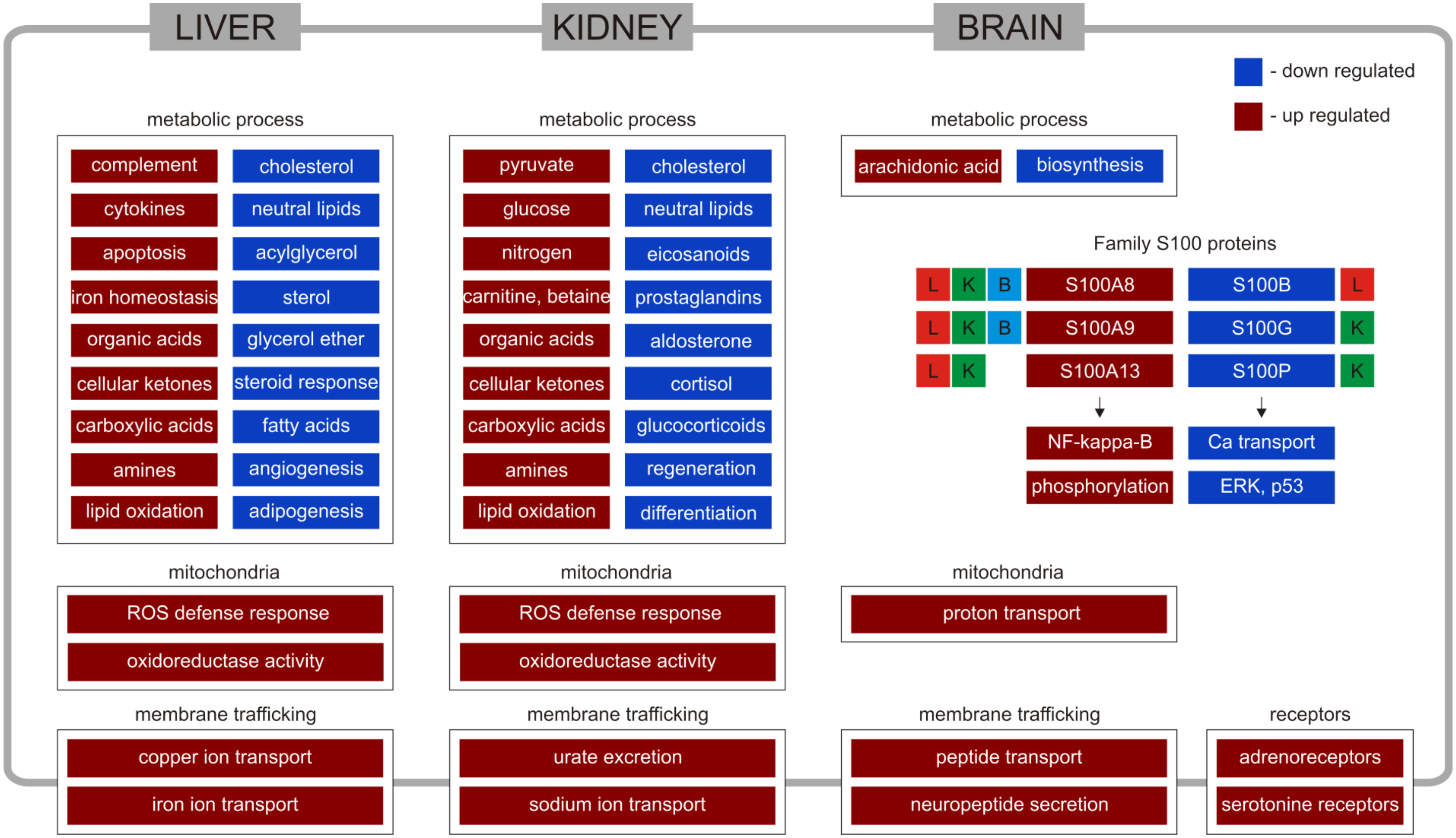
In the kidney, the protein products of differentially regulated genes were associated with metabolism of steroid hormones, and included downregulated cytochromes, aldo-keto reductase (Akr1d1) and steroidogenic acute regulatory protein (Star) (Supplementary Table 29). These proteins play key roles in progesterone, androstenedione, testosterone and pregnenolone metabolism (Supplementary Fig. 29). Thus, differential regulation of steroid biosynthesis may be an additional mechanism of adaptation to low oxygen. The most dramatic transcriptional changes in the NMR brain were associated with downregulation of protein biosynthesis and induction of synaptic transmission and neuropeptide signaling (Supplementary Table 30).

We also found reduction in transcription of carbonic anhydrase genes (*Ca*) 1, 2 and 3 in the liver, *Ca4* and *Ca8* in the brain, and *Ca11* in the kidney. In the liver, the expression of *Ca3* was reduced 300 fold under low O_2. At the same time, *Ca12, Ca13* and *Ca14* were upregulated in the kidney. *Ca8* and *Ca11* encode "acatalytic" CA isoforms whose function is unclear. Other CAs catalyze reversible hydration of CO_2 and are involved in maintaining the cellular pH. The differential regulation of CAs may be the mechanism allowing NMR to control CO_2 and bicarbonate concentrations. The concentration of CO_2 was constant in our model experiment. Thus, *Ca*s may be regulated by the same pathways controlling transcription in response to O_2 changes. Analysis of genes under positive selection further shed light on the understanding of tolerance to low O_2 as *Ca12* and phosphate-activated glutaminase (*Gls2)* genes were under rapid evolution.

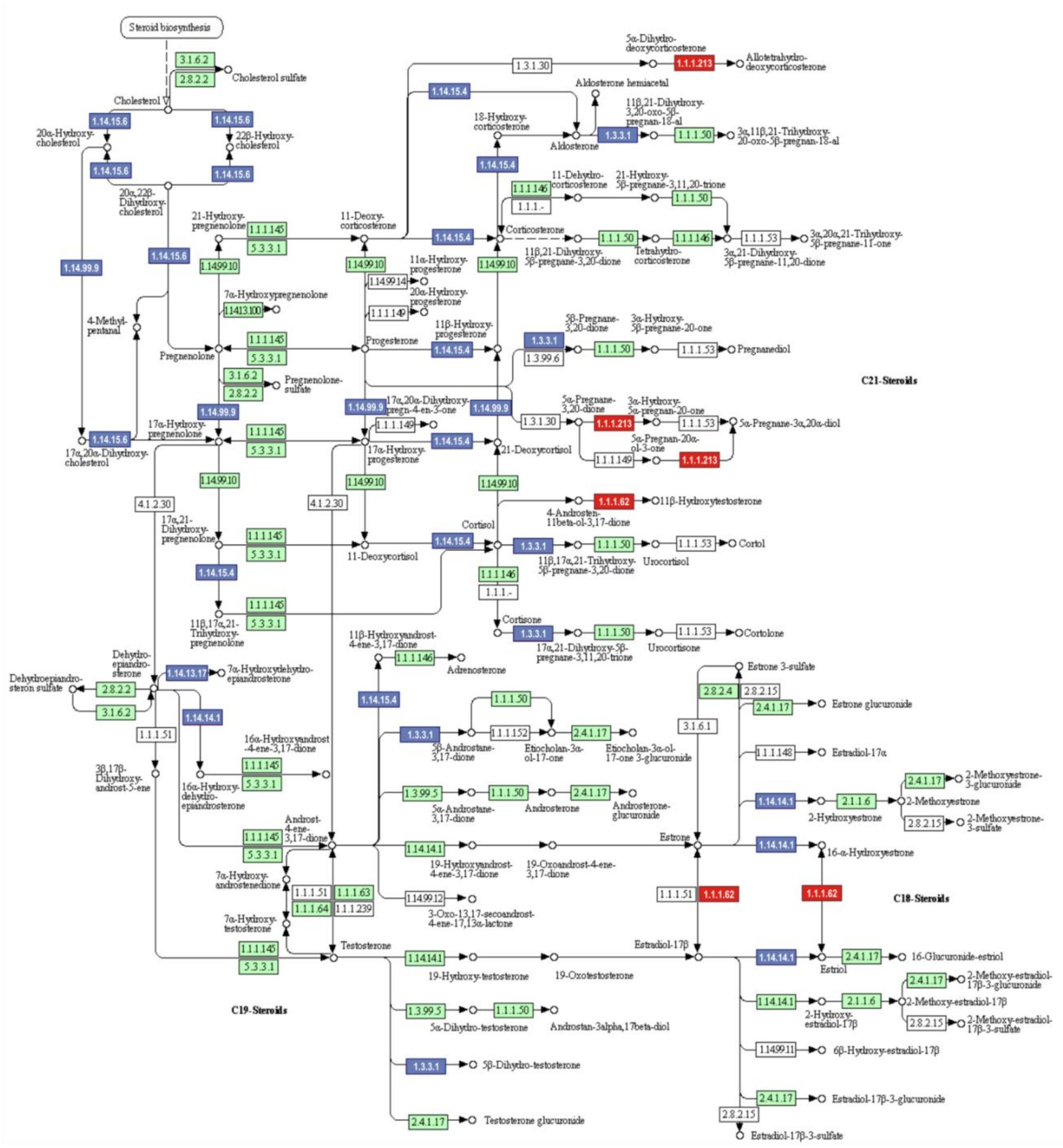
Our analysis showed that only a few genes were co-regulated in the three NMR tissues, arguing against a common mechanism of adaptation to low O_2 in them (Supplementary Fig. 27 and Supplementary Table 31). One of those was a Sox9-inducible growth factor (encoded by *Gdf10* gene), whose downregulation contrasted with the findings in humans. *Sox9* and *Sox9*-regulated genes are under positive transcription control of Hif1. Hif1 controls the expression of up to 2% of human genes and represents the major hypoxia protective mechanism in mammals. Hif1 is a heterodimer composed of alpha and beta subunits. The beta subunit has been identified as the aryl hydrocarbon receptor nuclear translocator (Arnt). We observed the induction of transcription of *Arnt2* in the kidney which encodes a paralog of Arnt. In addition, we found the induction of Hif1-interacting co-activators *NcoA1*, *Rora*, and *Hnf4* in the kidney and activators of *Hif1-α* transcription, *Ppara* and *AhR* (Supplementary Fig. 30). Taken together, these observations suggest the involvement of HIF1-mediated signaling pathways in NMR adaptation to low O_2 environment.



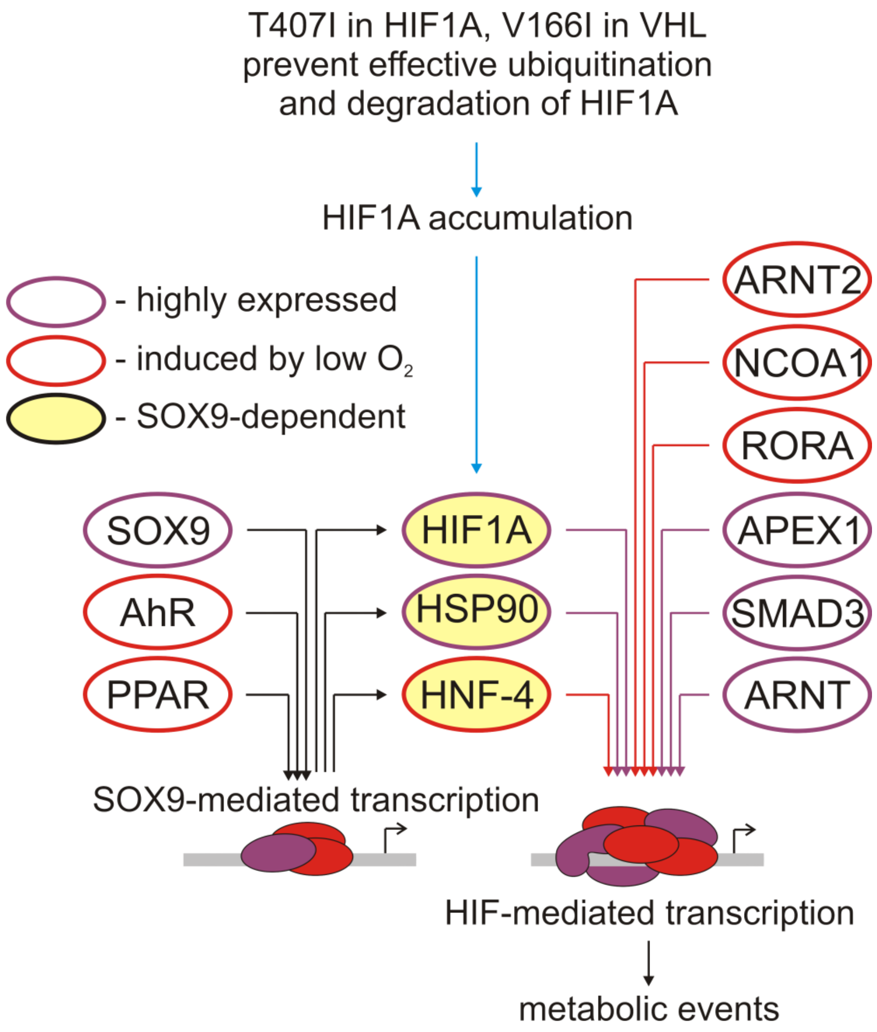
Supplementary Fig. 27. Venn diagrams showing the intersection of (A) down regulated and (B) up regulated genes in liver, kidney and brain of NMR subjected to low oxygen atmosphere.



Supplementary Fig. 28. Schematic diagram showing the differential nature of cellular processes occurring in liver, kidney and brain of NMR subjected to low oxygen. Up regulated events are shown in red, and down regulated events in blue. We observed differential regulation of S100 Ca^{2+} family proteins. S100A8 and S100A9 may up-regulate transcription of genes that are under the control of NF-kappa-B. S100A13 is required for the copper-dependent stress-induced export of IL1A and FGF1. S100B is involved in activation of STK38 kinase that is a negative regulator of MAP3K1/2 signaling. S100P is involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100G is vitamin D-dependent and its expression correlates with calcium transport activity. The role of these proteins in hypoxia currently is not known. Colored L, K, B – liver, kidney and brain, respectively.



Supplementary Fig. 29. A KEGG pathway diagram showing changes in steroid hormone biosynthesis in the kidney of NMR subjected to low oxygen. Down regulated genes (in blue): AKR1D1 (EC:1.3.1.3), CYP11A (EC:1.14.15.6), CYP17A1 (EC: 1.14.99.9), CYP11B1 (EC:1.14.15.4), CYP11B2 (EC:1.14.15.5), CYP1A1 (EC:1.14.14.1), CYP7A1 (EC:1.14.17.13). Up regulated genes (in red): HSD17B7 (EC:1.1.1.62), AKR1C1 (EC:1.1.1.213).



Supplementary Fig. 30. Putative cellular signaling events activated in response to the low oxygen conditions. SOX9, HIF1A, HSP90, APEX1, SMAD3 and ARNT are naturally overexpressed in NMR in comparison to mouse (rose ellipses). The transcription of HIF1A, HSP90 and HNF-4 can be activated by combination of SOX9, AhR and PPAR factors. The expression of AhR, PPAR, ARNT2, NCOA1 and RORA is induced during the ischemia (red ellipses). HIF1 was previously proposed as a primarily oxygen sensor. HIF1 (heterodimer of HIF1A and ARNT/ARNT2) is stabilized under the hypoxia leading to the formation of a functional transcription factor complex. This complex, in combination with transcriptional co-activators (shown on the right), is the master regulator of O_2 homeostasis and can induce a network of genes involved in angiogenesis, erythropoiesis, and glucose metabolism. Under normoxic conditions, HIF1A is hydroxylated by prolyl hydroxylases. This event leads to the recruitment of the pVHL E3 ligase complex to HIF1A. The pVHL E3 ligase complex ubiquitylates HIF1A, leading to its degradation. In NMR, T407I in HIF1A and V166I in VHL may prevent ubiquitin-dependent degradation of HIF1A resulting in accumulation of this protein in the cell.

8 Supplementary references

1 Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317, (2010).

2 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

3 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).

4 Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, Unit 4 (2004).

5 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

6 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995 (2004).

7 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** Suppl 2, ii215-225 (2003).

8 Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-522 (2000).

9 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

10 Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45-48 (2000).

11 Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* **396**, 59-70 (2007).

12 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

13 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

14 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715 (2004).

15 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007 (2002).

16 Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

17 Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572-580 (2006).

18 Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818 (1998).

19 Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).

20 Wikstrom, N., Savolainen, V. & Chase, M. W. Evolution of the angiosperms: calibrating the family tree. *Proc Biol Sci* **268**, 2211-2220 (2001).

21 Wang, Z. Q. A new Permian gnetalean cone as fossil evidence for supporting current molecular phylogeny. *Ann Bot* **94**, 281-288 (2004).

22 Yang, Z. PAML: a program package for phylogenetic analyses by maximum

likelihood. *Comp Appl BioSci* **13**, 555-556 (1997).

23 Yang, Z. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**,1586-1591 (2007).

24 Yang, Z., & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908-917 (2002).

25 Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol Biol Evol* **22**, 2472-2479 (2005).

26 Fletcher, W. and Yang, Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Mol Biol Evol* **27**, 2257-2267 (2010).

27 Loytynoja, A. & Goldman, N. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science* **320**, 1632-1635 (2008).

28 Markova-Raina, P. & Petrov, D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res* **21**, 863-874 (2011).

29 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).

30 Talavera, G., & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577 (2007).

31 Higgins, M.E. *et al*. CancerGenes: a gene selection resource for cancer genome projects. *Nucl Acids Res* **35**, D721-D726 (2007).

32 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).

33 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).

34 Chen, S. *et al.* De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS One* **5**, e15633 (2010).

35 Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Res* **7**, 986-995 (1997).

36 Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology* **10**, R130.1 (2009).